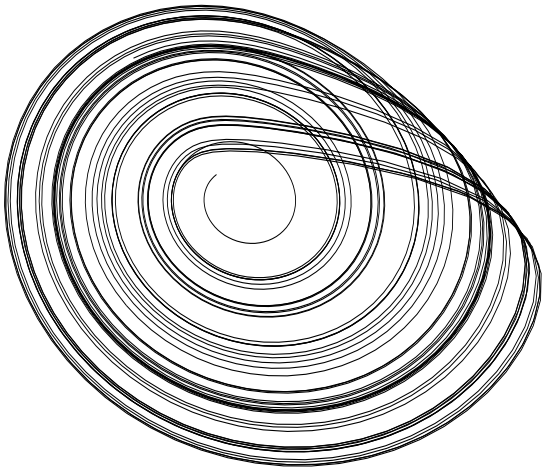
NONLINEAR DIFFERENTIAL EQUATIONS AND CHAOS

(AM303b Notes)



P. Yu and C. Essex

Department of Applied Mathematics
The University of Western Ontario
London, Ontario, Canada N6A 5B7

January 2003

Chapter 1

INTRODUCTION

This course is about ***nonlinear differential equations*** (primarily ordinary), but it is also about a change in scientific perspective that has occurred in the last 30 years or so. Although there are many interesting consequences and perspectives to this change, it primarily affects our views about cause and effect (determinism) and ultimately what predicting the future means in this area.

Had anyone predicted that new discoveries could be made in dynamics 300 years after the publication of Newton's Principia, they would have been thought naive or foolish. Yet, in the past three decades, new phenomena in nonlinear dynamics were discovered, principal among these being chaotic and unpredictable behavior from apparently deterministic systems.

This course will touch on both classical (algebraic, e.g. perturbation) methods and modern (geometric, e.g. phase-space) methods, but the emphasis will be put on the new geometrical ideas that are revolutionizing dynamical systems theory. For being an applied mathematician, an engineer, a scientist, an analyst or an experimentalist of all disciplines, we are usually concerned with modeling and understanding the time evolution of real systems. Most real-world problems confronting us, being neither linear nor even nearly linear, fall outside the domain of traditional closed-form analysis, and must be tackled in the first instance on a computer. But numerical simulations, like physical experiments, typically produce unwieldy masses of data, and for both of these, phase-space concepts must be recognized as an essential guide to the structuring of the investigation and the interpretation of the results.

At the present time we are indeed witnessing a spectacular blossoming of nonlinear dynamics, made possible on the one hand by great theoretical strides in Poincaré's qualitative topological approach and on the other hand by the wide availability of powerful computers. This has been stimulated and sustained by important and exciting new applications that have

multiplied throughout the biological, ecological, social and economic sciences, far beyond the still vibrant traditional fields of mechanics, physics, chemistry and engineering. Indeed, wherever the time evolution of a natural or human-made system needs to be modeled and explored, the subtle and versatile techniques of dynamical systems theory are being invoked.

A significant element of this major thrust, which nicely epitomizes its most advanced theoretical, computational, and experimental features, is the discovery and delineation of chaotic motions in remarkably simple deterministic models from a galaxy of disciplines ranging from population dynamics and meteorology to lasers and particle accelerators. The recent discovery of quite large regimes of chaos in the long-familiar, sinusoidally driven Duffing and Van der Pol oscillators emphasizes just how much was missed by the classical analysis of ordinary differential equations.

Before we start, we should agree about something: chaos is part of an even grander subject known as ***dynamics***. This is the subject which deals with change, with systems which evolve in time. Whether the system in question settles down in equilibrium, keeps repeating in cycles, or does something more complicated, it is **dynamics** that we use to analyze the behavior.

1.1 Brief History of Dynamics

The subject of dynamics began in the mid-1600s, when Newton invented differential equations, discovered his laws of motion and universal gravitation, and combined them to explain Kepler's laws of planetary motion. Specifically, Newton solved the two-body problem – the problem of calculating the motion of the earth around the sun. However, it was found later that Newton's analytical methods could not be extended to solve three-body problems, in the sense of obtaining explicit formulas for the motions of three interacting bodies. The breakthrough came with the work of Poincaré in the late 1800s. He developed a powerful *geometric* approach to analyzing such questions, emphasizing qualitative rather than quantitative properties.

In the first half of this century, dynamics was largely concerned with nonlinear oscillators and their applications in physics and engineering. Nonlinear oscillators played a vital role

in the development of such techniques as radio, radar, phase-locked loops, and lasers. The invention of the high-speed computer in the 1950s finally led to the meteorologist – Lorenz's discovery in 1963 of chaotic motion on a strange attractor. He studied a simplified model of convection rolls in the atmosphere and found that the solutions to his equations never settle down to equilibrium or to a periodic state – instead they continued to oscillate in an irregular, aperiodic fashion. Moreover, if he started his simulations from two slightly different initial conditions, the resulting behaviors would soon become totally different. This is known as being *sensitive* to the initial conditions of chaos.

Table 1.1 Dynamics – A Brief History

1666	Newton	Invention of calculus, explanation of planetary motion
1700s		Flowering of calculus and classical mechanics
1800s		Analytical studies of planetary motion
1890s	Poincaré	Geometric approach, nightmares of chaos
1920–1950		Nonlinear oscillators in physics and engineering, invention of radio, radar, laser
1920–1960	Birkhoff Kolmogorov Arnol'd Moser	Complex behavior in Hamiltonian mechanics
1963	Lorenz	Strange attractors in simple model of convection
1970s	Ruelle & Takens	Turbulence and chaos
	May	Chaos in logistic map
	Feigenbaum	Universality and renormalization, connection between chaos and phase transition
		Experimental studies of chaos
	Winfree	Nonlinear oscillators in biology
	Mandelbrot	Fractals
1980s		Widespread interest in chaos, fractals, oscillators, and their applications

Lorenz's work had little impact until the 1970s, the boom years for chaos. Here are some of the main developments of that glorious decade. In 1971 Ruelle and Takens proposed a new theory for the onset of turbulence in fluids, based on abstract considerations about strange attractors. A few years later, May found examples of chaos in iterated mappings arising in population biology. Next the physicist Feigenbaum surprisingly discovered that there are certain universal laws governing the transition from regular to chaotic behavior; roughly speaking, completely different systems can go chaotic in the same way. At the same time, many experimentalists tested the new ideas about chaos in experiments on fluids, chemical reactions, electronic circuits, mechanical oscillators, and semiconductors. There were two other major developments in dynamics in the 1970s. Mandelbrot codified and popularized fractals, produced magnificent computer graphics of them, and showed how they could be applied in a variety of subjects. And in the emerging area of mathematical biology, Winfree applied the geometric methods of dynamics to biological oscillations. By the 1980s many people were working on dynamics, with contributions too numerous to list. Table 1.1 summarizes this history.

1.2 An Overview of Nonlinear Phenomena

There are two main types of dynamical systems: ***differential equations*** and ***iterated maps*** (also known as difference equations). Differential equations describe the evolution of systems in continuous time, whereas iterated maps arise in problems where time is discrete. For differential equations, the main distinction is between ordinary and partial differential equations. For instance, the equation describing the motion of a real pendulum is given by

$$\frac{d^2x}{dt^2} + \frac{g}{l}\sin x = 0 \tag{1.1}$$

where x is the angle of the pendulum from vertical, g is the acceleration due to gravity, and l is the length of the pendulum. This equation is an ordinary differential equation, because it involves only the ordinary derivative $\dfrac{d^2x}{dt^2}$. That is, there is only one independent variable, the time t. In contrast, the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{1.2}$$

is a partial differential equation – it has both time t and space x as independent variables. The main attention of this course will be focused on ordinary differential equations and difference equations.

A general form for ordinary differential equations can be described by the system

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, \cdots, x_n) \\ \dot{x}_2 &= f_2(x_1, x_2, \cdots, x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, x_2, \cdots, x_n) \end{aligned} \tag{1.3}$$

where the overdots denote differentiation with respect to t, i.e. $\dot{x}_i \equiv \dfrac{dx_i}{dt}$. The variables $x_1, x_2, \cdots, x_n$ are called state variables, and the functions $f_1, f_2, \cdots, f_n$ are determined by the properties of the problem. The oscillator (1.1) can be rewritten in form of (1.3) as

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -\frac{g}{l}\sin x_1 \end{aligned} \tag{1.4}$$

by introducing $x_1 = x$ and $x_2 = \dot{x}$. This equation is ***nonlinear***. Nonlinearity makes the pendulum equation very difficult to solve analytically. However, for small x, since $\sin x \approx x$, equation (1.4) can be approximated by

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -\frac{g}{l}\, x_1 \end{aligned} \tag{1.5}$$

which is a ***linear*** equation and can then be solved easily. But by restricting to small x, we are throwing out some of the physics, like motions where the pendulum whirls over the top.

Ed Lorenz (1963) derived a three-dimensional chaotic system from a drastically simplified model of convection rolls in the atmosphere. This well-known equation is now called Lorenz equations, given by

$$\begin{aligned} \dot{x} &= \sigma\,(y - x) \\ \dot{y} &= \rho\, x - y - x\, z \\ \dot{z} &= x\, y - \beta\, z \end{aligned} \tag{1.6}$$

where σ, ρ, $\beta > 0$ are parameters. The same equations also arise in models of lasers and dynamos.

Lorenz discovered that this simple-looking deterministic system could have extremely erratic dynamics: over a wide range of parameters, the solutions oscillate irregularly, never exactly repeating but always remaining in a bounded region that they settled onto a complicated set, now called a *strange attractor* (see Fig. 1.1).

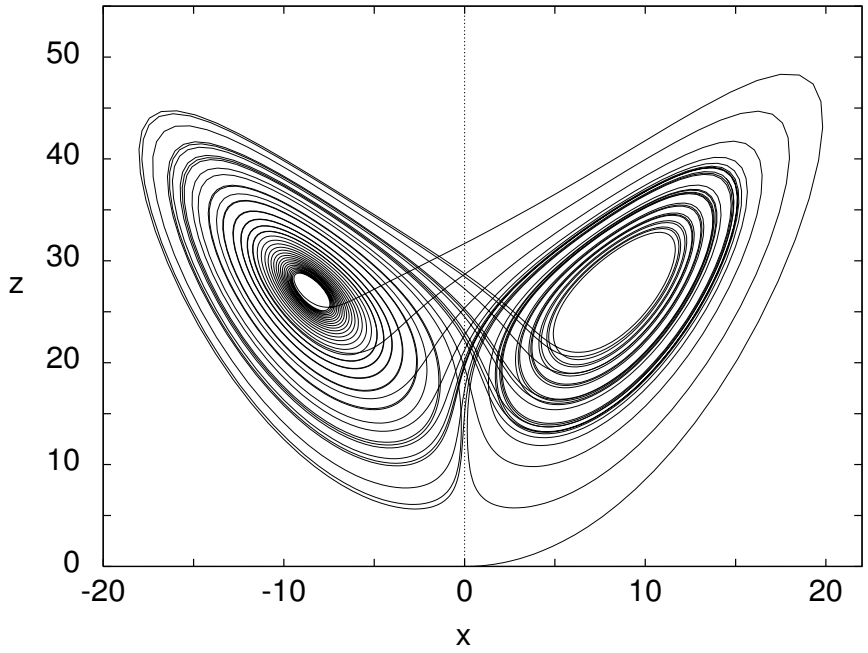


Fig. 1.1 Strange attractor of the Lorenz system.

1.3 A Dynamical View of the World

Having established the ideas of nonlinearity and phase space, we can now present a framework for dynamics and its applications, which is shown in Fig. 1.2. The framework has two axes. One axis tells us the number of variables needed to characterize the state of the system, i.e. the dimension of the phase space. The other axis tells us whether the system is linear or nonlinear. For example, the equation describing the exponential growth

of a population of organisms is given by the first-order differential equation,

$$\dot{x} = \mu\, x$$

where x is the population at time t and μ is the growth rate. We place this system in the column labeled "$n = 1$" because *one* piece of information – the current value of the population – is sufficient to predict the population at any later time. The system is also classified as linear. As a second example, the equation of the pendulum (1.1) or (1.4) is given by *two* variables: its current angle x and angular velocity $\dot{x}$. Because two variables are needed to specify the state, the pendulum belongs in the $n = 2$ column of Fig. 1.2. Moreover, the system is nonlinear. Hence the pendulum is in the lower, nonlinear half of the $n = 2$ column.

One can continue to classify systems in this way, and the result will be something like the framework shown in Fig. 1.2. Admittedly, some aspects of the picture are debatable. You might think that some topics should be added, or placed differently, or even that more axes are needed – the point is to think about classifying systems on the basis of their dynamics.

There are some striking patterns in Fig. 1.2. All of the simplest systems occur in the upper left-hand corner. These are the small linear systems. Roughly speaking, these linear systems exhibit growth, decay, or equilibrium when $n = 1$, or oscillations when $n = 2$. The italicized phrases in this figure indicate that these broad classes of phenomena first arise in this part of the diagram.

The next most familiar part of the picture is the upper right-hand corner. This is the domain of classical applied mathematics and mathematical physics where the linear partial differential equations live. These partial differential equations involve an infinite "continuum" of variables because each point in space contributes additional degrees of freedom. Even though these systems are large, they are tractable by using linear techniques such as Fourier analysis and transform methods.

The lower half of Fig. 1.2 is the nonlinear part. As we increase the phase space dimension from $n = 1$ to $n = 3$, we encounter new phenomena at every step, from fixed points and bifurcations when $n = 1$, to nonlinear oscillations when $n = 2$, and finally chaos and

fractals when $n = 3$. In all cases, a geometric approach proves to be very powerful, and gives us most of the information we want, even though we usually can't solve the equations in the traditional sense of finding a formula for the answer. Our journey will also take us to some of the most exciting parts of modern science, such as mathematical biology and condensed-matter physics.

It is noted that the framework also contains a region forbiddingly marked "The frontier". These topics are not completely unexplored, of course, but it is fair to say that they lie at the limits of current understanding. The problems are very hard, because they are both large and nonlinear. The resulting behavior is typically complicated in *both space and time*, as in the motion of a turbulent fluid or the patterns of electrical activity in a fibrillating heart.

Number of variables →

Nonlinearity: Linear ↓ Nonlinear

n = 1	n = 2	n > 2	n > > 1	Continuum
Growth, decay, or equilibrium	***Oscillations***		***Collective phenomena***	***Waves and patters***
Exponential growth	Linear oscillator	Civil engineering structures	Coupled harmonic oscillators	Elasticity
RC circuit	Mass and spring	Electrical engineering	Solid-state physics	Wave equations
Radioactive decay	RLC circuit		Molecular dynamics	Electromagnetism (Maxwell)
	2-body problem (Kepler, Newton)		Equilibrium statistical Mechanics	Quantum mechanics (Schrodinger, Heisenberg, Dirac)
				Heat and diffusion
				Acoustics
				Viscous fluids
		The frontier		
		Chaos		***Spatio-temporal complexity***
Fixed points	Pendulum	Stranger attractors (Lorenz)	Coupled nonlinear oscillators	Nonlinear waves (shocks, solitons)
Bifurcation	Anharmonic oscillators	3-body problem (Poincare)	Lasers, nonlinear optics	plasmas
Overdamped systems relaxational dynamics	Limit cycles	Chemical kinetics	Nonequilibrium statistical mechanics	Earthquakes
Logistic equation for single species	Biological oscillators (neurons, heart cells)	Iterated maps (Feigenbaum)	Nonlinear solid-state physics (semiconductors)	General relativity (Einstein)
	Predator-prey cycles	Fractals (Mandelbrot)	Josephson arrays	Quantum field theory
	Nonlinear electronics (Van der Pol, Josephson)	Forced nonlinear oscillators (Levinson, Smale)	Heart cell synchronization	Reaction-diffusion, biological and chemical waves
		Practical uses of chaos	Neural networks	Fibrillation
		Quantum chaos ?	Immune system	Epilepsy
			Ecosystems	Turbulent fluids (Navier-Stokes)
			Economics	Life

Fig. 1.2 A dynamical view of the world

Chapter 2

BRIEF REVIEW OF LINEAR ODE

In this chapter we will give a brief review of linear ordinary differential equations (ODE). The main attention will focus on the systems with constant coefficients. (General theory of linear differential equations may be found in, for example, "Advanced Engineering Mathematics" by Peter V. O'Neil, 1991.) These linear ODE are important because many physical and engineering problems can be described by these kind of ODE. On one hand, we need to find solutions and study the properties of the solutions; on the other hand, many ODE can be approximated by first-order ODE.

2.1 Homogeneous Differential Equations

We start with the nth-order differential equation, given by

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \cdots + a_n y = 0 \tag{2.1}$$

where a_i's are constants (real or complex). We will also use prime to denote differentiations, thus, y', y'' $\cdots$ $y(n)$, etc. represent the 1st, 2nd, $\cdots$ n-th derivatives of y. Use the differential operator $D = \dfrac{d}{dt}$, one obtains

$$\frac{d^2}{dt^2} = D^2\,, \quad \frac{d^3}{dt^3} = D^3\,, \quad \cdots \quad \frac{d^n}{dt^n} = D^n \tag{2.2}$$

and define the operator function

$$L(D) = D^n + a_1 D^{n-1} + a_2 D^{n-2} + \cdots + a_n\,, \tag{2.3}$$

then,

$$L(D)\, y = D^n y + a_1 D^{n-1} y + a_2 D^{n-2} y + \cdots + a_n y\,. \tag{2.4}$$

Therefore, the linear DE (2.1) can be written as

$$L(D)\, y = 0\,. \tag{2.5}$$

Suppose that $L(D)$ and $M(D)$ are two polynomials of D, then

$$\begin{aligned} L(D)\,(y_1 + y_2) &= L(D)\,y_1 + L(D)\,y_2 \\ [\,L(D) + M(D)\,]\,y &= L(D)\,y + M(D)\,y \\ L(D)\,[\,M(D)\,y\,] &= [\,L(D)\,M(D)\,]\,y\,. \end{aligned} \tag{2.6}$$

Moreover,

$$\begin{aligned} D\,e^{\lambda t} &= \frac{d}{dt}\left(e^{\lambda t}\right) = \lambda\,e^{\lambda t} \\ D^2\,e^{\lambda t} &= \lambda^2\,e^{\lambda t} \\ &\vdots \\ D^n\,e^{\lambda t} &= \lambda^n\,e^{\lambda t} \end{aligned} \tag{2.7}$$

which indicates that

$$L(D)\,e^{\lambda t} = L(\lambda)\,e^{\lambda t}. \tag{2.8}$$

Hence, if λ is a root of the polynomial $L(\lambda)$, then function $e^{\lambda t}$ is a solution of the DE (2.1). If $e^{\lambda t}$ is a solution of (2.1), then λ must be the root of $L(D)$. We call

$$L(\lambda) = 0 \tag{2.9}$$

the *characteristic function* or *characteristic polynomial* (C.P.) of the DE (2.1).

2.1.1 Distinct roots. If the C.P. has n different roots $\lambda_1\,,\ \lambda_2\,,\ \cdots\,,\ \lambda_n\,$, then it can be shown that the n functions

$$y_i(t) = e^{\lambda_i t}\,, \quad i = 1,\ 2,\ \cdots n \tag{2.10}$$

are independent and construct the basic solutions of the DE (2.1). Thus, the general solution of (2.1) can be given as a linear combination of these basic solutions:

$$y(t) = \sum_{i=1}^{n} C_i\,e^{\lambda_i t} \tag{2.11}$$

where $C_i\,,\ i = 1,\ 2,\ \cdots\,,\ n$ are arbitrary constants.

When all coefficients a_i in (2.1) are real and if the C.P. (2.9) has a complex root $\lambda = \mu + i\,\nu$, then the complex conjugate $\bar{\lambda} = \mu - i\,\nu$ is also the root of $L(\lambda)$. Thus, both

$$e^{\lambda t} = e^{(\mu+i\nu)t} = e^{\mu t}\, e^{i\nu t} = e^{\mu t}\,(\cos\nu t + i\,\sin\nu t) \tag{2.12}$$

(where *Euler's formula* has been used) and

$$e^{\bar{\lambda} t} = e^{\mu t}\,(\cos\nu t - i\,\sin\nu t) \tag{2.13}$$

are solutions of the DE (2.1). So, the real functions

$$y = \frac{1}{2}\left(e^{\lambda t} + e^{\bar{\lambda} t}\right) = e^{\mu t}\,\cos\nu t \tag{2.14}$$

and

$$y = \frac{1}{2\,i}\left(e^{\lambda t} - e^{\bar{\lambda} t}\right) = e^{\mu t}\,\sin\nu t \tag{2.15}$$

are solutions of (2.1).

In general, suppose $L(\lambda) = 0$ has $2\,k$ complex roots

$$\begin{aligned} \lambda_1 &= \bar{\lambda}_2 = \mu_1 + i\,\nu_1 \\ \lambda_2 &= \bar{\lambda}_4 = \mu_2 + i\,\nu_2 \\ &\vdots \\ \lambda_{2k-1} &= \bar{\lambda}_{2k} = \mu_k + i\,\nu_k\,, \end{aligned} \tag{2.16}$$

and $n-2\,k$ real roots $\lambda_{2k+1}\,,\ \lambda_{2k+2}\,,\ \cdots\ \lambda_n\,$, then, $e^{\mu_1 t}\,\cos\nu_1 t\,,\ e^{\mu_1 t}\,\sin\nu_1 t\ \cdots,\ e^{\mu_k t}\,\cos\nu_k t\,,$ $e^{\mu_k t}\,\sin\nu_k t\,,\ e^{\lambda_{2k+1} t}\ \cdots,\ e^{\lambda_n t}$ are basic solutions of the DE (2.1), each of which is a reals function.

EXAMPLE 2.1. Find all solutions of the DE

$$y'' + 3\,y' + 2\,y = 0 \qquad y \in \boldsymbol{R}\,. \tag{2.17}$$

Solution: The C.P. of (2.17) is

$$L(\lambda) = \lambda^2 + 3\,\lambda + 2 = (\lambda + 1)\,(\lambda + 2) \tag{2.18}$$

which yields roots $\lambda_1 = -1$ and $\lambda_2 = -2$. So, the general solution of (2.17) is

$$y(t) = C_1\, e^{-t} + C_2\, e^{-2t} \tag{2.19}$$

where C_1 and C_2 are arbitrary real constants. (If $y \in \boldsymbol{C}$ then C_1 and C_2 are complex constants.)

EXAMPLE 2.2. Find all solutions of the DE

$$\frac{d^3y}{dt^3} - 3\,\frac{d^2y}{dt^2} + 9\,\frac{dy}{dt} + 13\,y = 0 \qquad y \in \boldsymbol{R}\,. \tag{2.20}$$

Solution: The characteristic polynomial is

$$L(\lambda) = \lambda^3 - 3\,\lambda^2 + 9\,\lambda + 13 = (\lambda + 1)\,(\lambda^2 - 4\,\lambda + 13) \tag{2.21}$$

which, in turn, results in $\lambda_1 = -1,\ \lambda_2 = 2+3\,i$ and $\lambda_2 = 2-3\,i$. Hence the general solution of (2.20) is given by

$$y(t) = C_1\, e^{-t} + C_2\, e^{2\,t}\, \cos 3t + C_3\, e^{2t}\, \sin 3t \tag{2.22}$$

where $C_i\ = 1,\ 2,\ 3$ are arbitrary real constants.

The complex form solution of (2.20) is

$$y(t) = C_1\, e^{-t} + C_2\, e^{(2+3i)t} + \bar{C}_2\, e^{(2-3i)t} \tag{2.23}$$

where C_1 is a real constant while C_2 is a complex constant.

EXAMPLE 2.3. Find all solutions of damped system

$$m\,\ddot{x} + k\,x = 0 \qquad (m,\ k > 0). \tag{2.24}$$

Solution: Note that we also frequently use dot to denote differentiations with respect to time. (If the independent variable is not time, then we usually use prime, not the dot, to denote the differentiations.) The C.P. of (2.24) is

$$L(\lambda) = m\,\lambda^2 + k = 0 \qquad \Longrightarrow \qquad \lambda_{\pm} = \pm\sqrt{\frac{k}{m}}\; i\,. \tag{2.25}$$

Thus,

$$x(t) = C_1\, \cos\sqrt{\frac{k}{m}}\; t \;+\; C_2\, \sin\sqrt{\frac{k}{m}}\; t \tag{2.26}$$

where C_1 and C_2 are arbitrary real constants. Suppose at time $t = 0$,

$$x(0) = x_0 \qquad \text{and} \qquad \dot{x}(0) = \dot{x}_0\,, \tag{2.27}$$

then

$$x_0 = x(0) = C_1 \qquad \text{and} \qquad \dot{x}_0 = \dot{x}(0) = C_2\sqrt{\frac{k}{m}} \quad \Longrightarrow \quad C_2 = \sqrt{\frac{m}{k}}\,\dot{x}_0\,. \tag{2.28}$$

Hence,

$$x(t) = x_0\,\cos\sqrt{\frac{k}{m}}\,t \;+\; \dot{x}_0\sqrt{\frac{m}{k}}\,\sin\sqrt{\frac{k}{m}}\,t \tag{2.29}$$

is a solution satisfying the initial condition (2.27).

Let $\omega = \sqrt{\dfrac{k}{m}}$ which is called *natural frequency* of system (2.24), and (2.24) becomes

$$\ddot{x} + \omega^2\,x = 0. \tag{2.30}$$

The solution of (2.26) can now be written as

$$x(t) = C_1\,\cos\omega t + C_2\,\sin\omega t = r\,\cos(\omega t + \phi) \tag{2.31}$$

where

$$r = \sqrt{C_1^2 + C_2^2} \qquad \text{and} \qquad \phi = -\tan^{-1}\left(\frac{C_2}{C_1}\right). \tag{2.32}$$

2.1.2 Repeated roots. Next, let's discuss when $L(\lambda)$ has repeated roots. We know that if λ_1 and λ_2 are the roots of $L(\lambda) = 0$, then $\dfrac{1}{\lambda_1 - \lambda_2}\left(e^{\lambda_1 t} - e^{\lambda_2 t}\right)$ is also the root of $L(\lambda) = 0$. But

$$\begin{aligned}
&\lim_{\lambda_2\to\lambda_1}\frac{1}{\lambda_1-\lambda_2}\left(e^{\lambda_1 t}-e^{\lambda_2 t}\right)\\
=\;&\lim_{\lambda_2\to\lambda_1}\frac{1}{\lambda_1-\lambda_2}e^{\lambda_1 t}\left[1-e^{(\lambda_2-\lambda_1)t}\right]\\
=\;&e^{\lambda_1 t}\lim_{s\to 0}\frac{1-e^{st}}{-s}\qquad (s=\lambda_2-\lambda_1)\\
=\;&e^{\lambda_1 t}\,t\lim_{s\to 0}e^{st}\qquad (\text{L'Hospital Rule})\\
=\;&t\,e^{\lambda_1 t}.
\end{aligned} \tag{2.33}$$

This indicates that if λ_1 is a twice repeated root of $L(\lambda) = 0$, then $e^{\lambda_1 t}$ and $t\,e^{\lambda_1 t}$ are all solutions of (2.1). Similarly, if λ_1 is a k time repeated root of $L(\lambda) = 0$, then $e^{\lambda_1 t},\ t\,e^{\lambda_1 t},\ \cdots\ t^{k-1}\,e^{\lambda_1 t}$ are all solutions of (2.1).

Suppose $\lambda_1\,,\ \lambda_2\,,\ \cdots\ \lambda_m$ are distinct roots of $L(\lambda) = 0$, and they repeat, respectively, $k_1\,,\ k_2\,,\ \cdots\ k_m$ times, such that $k_1 + k_2 + \cdots k_m = n$, then the functions

$$\begin{aligned}
y_1(t) &= e^{\lambda_1 t}, \qquad y_2(t) = t\,e^{\lambda_1 t}, \quad \cdots \quad y_{k_1}(t) = t^{k_1-1}\,e^{\lambda_1 t},\\
y_{k_1+1}(t) &= e^{\lambda_2 t}, \qquad y_{k_1+2}(t) = t\,e^{\lambda_2 t}, \quad \cdots \quad y_{k_1+k_2}(t) = t^{k_2-1}\,e^{\lambda_2 t},\\
&\vdots\\
y_{n-k_m+1}(t) &= e^{\lambda_m t}, \qquad y_{n-k_m+2}(t) = t\,e^{\lambda_m t}, \quad \cdots \quad y_n(t) = t^{k_m-1}\,e^{\lambda_m t}
\end{aligned}$$

are the *basic solutions* of the DE (2.1).

EXAMPLE 2.4. Find all solutions of the DE

$$\frac{d^3y}{dt^3} + 3\,\frac{d^2y}{dt^2} + 3\,\frac{dy}{dt} + y = 0. \tag{2.34}$$

Solution: The C.P. of (2.34) is

$$L(\lambda) = \lambda^3 + 3\,\lambda^2 + 3\,\lambda + 1 = (\lambda + 1)^3. \tag{2.35}$$

So $\lambda = -1$ is a repeated roots (3 times) of (2.34). Hence,

$$y(t) = (C_1 + C_2\,t + C_3\,t^2)\,e^{-t} \tag{2.36}$$

is the general solution, in which C_i are arbitrary constants.

EXAMPLE 2.5. Find all solutions of the DE

$$\frac{d^4y}{dt^4} + 2\,\frac{d^2y}{dt^2} + y = 0. \tag{2.37}$$

Solution:

$$L(\lambda) = \lambda^4 + 2\,\lambda^2 + 1 = (\lambda^2 + 1)^2 \tag{2.38}$$

which yields $\lambda_1 = i$ (repeated twice) and $\lambda_2 = -\,i$ (repeated twice). Thus

$$y(t) = (C_1 + C_2 t)\,e^{it} + (C_3 + C_4 t)\,e^{-it} \tag{2.39}$$

where C_i are complex constants. The real solution can be found following the procedure of single roots. We conclude that

$\cos t, \quad \sin t, \quad t\cos t \quad \text{and} \quad t\sin t$

are the basic solutions of the DE (2.37). Hence,

$$y(t) = (C_1 + C_2 t)\cos t + (C_3 + C_4 t)\sin t \tag{2.40}$$

where C_i are real constants.

2.2 Non-homogeneous Differential Equations

A homogeneous DE, as we discussed in the previous section, can be written in the general form

$$L(D)y = 0. \tag{2.41}$$

Non-homogeneous DEs are given by

$$L(D)y = f(t). \tag{2.42}$$

The solution of (2.42) can be found using the following *principle of superposition*:

(i) If $\phi^p(t)$ is a solution of (2.42) and $\phi^h(t)$ is a solution of (2.41), then $\phi(t) = \phi^h(t) + \phi^p(t)$ is a solution of (2.42).

(ii) If $\phi_1(t)$ and $\phi_2(t)$ are both solutions of (2.42), then $\phi(t) = \phi_1(t) + \phi_2(t)$ is also a solution of (2.42).

(iii) If $\phi(t)$ and $\psi(t)$ are solutions of the DE $L(D)y = f_1(t)$ and $L(D)y = f_2(t)$, respectively, then $\phi(t) = a\,\phi(t) + b\,\psi(t)$ is a solution of the DE $L(D)y = a\,f_1(t) + b\,f_2(t)$, where a and b are constants.

Now suppose that the basic solutions of the corresponding homogeneous DE of (2.42) are $\phi_1(t),\ \phi_2(t),\ \cdots\ \phi_n(t)$. Then the general solution of (2.42) is given by

$$y(t) = \sum_{i=1}^{n} C_k\,\phi_k(t) + \phi^p(t) \tag{2.43}$$

where C_k are arbitrary constants, and $\phi^p(t)$ is a *particular solution* of (2.42).

Having found the basic solutions of the corresponding homogeneous DE, a particular solution of (2.42) can be found by using methods such as the *method of undetermined coefficients* or *variation of parameters.*

2.3 The Method of Undetermined Coefficients

Suppose that the function $f(t)$ is given in so called *pseudo-polynomial* form:

$$f(t) = \sum_{j=1}^{k} P_j(t)\, e^{\alpha_j t}, \tag{2.44}$$

where $P_j(t)$ are polynomials of t, and α_j are constants. Since

$$\begin{aligned} e^{\mu t}\cos\nu t &= \frac{1}{2}\left[e^{(\mu+i\nu)t} + e^{(\mu-i\nu)t}\right] \\ e^{\mu t}\sin\nu t &= \frac{1}{2i}\left[e^{(\mu+i\nu)t} - e^{(\mu-i\nu)t}\right] \end{aligned}$$

so $e^{\mu t}\,[R(t)\cos\nu t + S(t)\sin\nu t]$ is also pseudo-polynomial, where $R(t)$ and $S(t)$ are polynomials of t.

According to the principle of superposition, if we find a solution for each term of $f(t)$, then adding them together results in a particular solution of the DE

$$L(D)y = f(t) \equiv \sum_{j=1}^{k} P_j(t)\, e^{\alpha_j t}.$$

Consider $f_j(t) = P_j(t)\, e^{\alpha_j t}$ then the DE

$$L(D)y = f_j(t) \equiv P_j(t)\, e^{\alpha_j t} \tag{2.45}$$

has a solution in the form of

$$y_j(t) = Q_j(t)\, e^{\alpha_j t}, \tag{2.46}$$

where $Q_j(t)$ is a polynomial of t, whose degree $\leq m + s$, m is the degree of $P_j(t)$ while

$$s = \begin{cases} 0 & \text{if } \alpha_j \text{ is not the roots of } L(\lambda) = 0 \quad (Non-resonant\ Case); \\ l & \text{if } \alpha_j \text{ is the roots of } L(\lambda) = 0, \text{ repeated } l \text{ times} \quad (Resonant\ case). \end{cases} \tag{2.47}$$

EXAMPLE 2.6. Find all solutions of the DE

$$\begin{aligned} \frac{dx}{dt} &= x + 2\,y - e^{-t} \\ \frac{dy}{dt} &= 4\,x + 3\,y + 4\,e^{-t}. \end{aligned} \tag{2.48}$$

First, we can obtain a 2nd order DE for x or y as follows: rewrite (2.48) in the form

$$\begin{aligned} D\,x - x - 2\,y &= -\,e^{-t} \\ -\,4\,x + D\,y - 3\,y &= 4\,e^{-t} \end{aligned} \tag{2.49}$$

or

$$\begin{aligned} (D-1)\,x - 2\,y &= -\,e^{-t} \\ -\,4\,x + (D-3)\,y &= 4\,e^{-t} \end{aligned}$$

which yields

$$\begin{aligned} [(D-1)\,(D-3) - 8]\;x &= \quad 12\,e^{-t} \\ [(D-1)\,(D-3) - 8]\;y &= -\,12\,e^{-t} \end{aligned}$$

by eliminating either x or y. So we have

$$\frac{d^2x}{dt^2} - 4\,\frac{dx}{dt} - 5\,x = 12\,e^{-t} \tag{2.50}$$

and so

$$L(\lambda) = \lambda^2 - 4\,\lambda - 5 \qquad \Longrightarrow \qquad \lambda_1 = -\,1 \quad \text{and} \quad \lambda_2 = 5. \tag{2.51}$$

Hence, the basic solutions of (2.50) are e^{-t} and e^{5t}. To find a particular solution of (2.50), notice that $f(t) = 12\,e^{-t}$ and $\alpha = -\,1$ which is a single root of the C.P. Since the degree of $P(t) = 12$ is 0 and $l = 1$, so $m + s = 1$. Therefore, the particular solution should be in the form of

$$x^p(t) = (a\,t + b)\,e^{-t}. \tag{2.52}$$

Differentiating $x^p(t)$ once and twice yields

$$\begin{aligned} \frac{dx^p}{dt} &= a\,e^{-t} - (a\,t + b)\,e^{-t} = -\,(a\,t + b - a)\,e^{-t} \\ \frac{d^2x^p}{dt^2} &= -\,a\,e^{-t} + (a\,t + b - a)\,e^{-t} = (a\,t + b - 2\,a)\,e^{-t}. \end{aligned} \tag{2.53}$$

Substituting (2.52) and (2.53) into (2.50) gives

$$(a\,t + 4\,a\,t - 5\,a\,t + b - 2\,a + 4\,b - 4\,a - 5\,b)\,e^{-t} = 12\,e^{-t} \tag{2.54}$$

which, in turn, results in $a = -2$ by balancing the coefficients. Thus, $x^p(t) = (b - 2\,t)\,e^{-t}$ and so the general solution is

$$x(t) = C_1\,e^{-t} + C_2\,e^{5t} + (b - 2\,t)\,e^{-t}. \tag{2.55}$$

Note that b has not be determined yet. Similarly, we may find

$$y(t) = C_3\,e^{-t} + C_4\,e^{5t} + (d + 2\,t)\,e^{-t}. \tag{2.56}$$

In order to determine the relations between the coefficients C_i, b and d, we substitute the solutions back to the original DE (2.48) to obtain

$$\frac{dx}{dt} - x - 2\,y + e^{-t} = 0$$

$$\Longrightarrow$$

$$-C_1\,e^{-t} + 5\,C_2\,e^{5t} + (2\,t - b - 2)\,e^{-t} - C_1\,e^{-t} - C_2\,e^{5t} - (b - 2\,t)\,e^{-t}$$
$$-2\,C_3\,e^{-t} - 2\,C_4\,e^{5t} - 2\,(d + 2\,t)\,e^{-t} + e^{-t} = 0$$

$$\Longrightarrow$$

$$-(2\,C_1 + 2\,C_3 + 2\,b + 2\,d + 1)\,e^{-t} + (4\,C_2 - 2\,C_4)\,e^{5t} = 0 \tag{2.57}$$

$$\Longrightarrow$$

$$-2\,C_1 - 2\,C_3 = 2\,b + 2\,d + 1 \qquad \text{and} \qquad C_4 = 2\,C_2\,. \tag{2.58}$$

Similarly, one has

$$\frac{dy}{dt} - 4\,x - 3\,y - 4\,e^{-t} = 0$$

$$\Longrightarrow$$

$$-C_3\,e^{-t} + 5\,C_4\,e^{5t} - (2\,t + d - 2)\,e^{-t} - 4\,C_1\,e^{-t} - 4\,C_2\,e^{5t} - 4\,(b - 2\,t)\,e^{-t}$$
$$-3\,C_3\,e^{-t} - 3\,C_4\,e^{5t} - 3\,(d + 2\,t)\,e^{-t} - 4\,e^{-t} = 0$$

$$\Longrightarrow$$

$$-(4\,C_1 + 4\,C_3 + 4\,b + 4\,d + 2)\,e^{-t} - (4\,C_2 - 2\,C_4)\,e^{5t} = 0 \tag{2.59}$$

which again yields the same equation (2.58). Let $C_3 = -C_1$, then $2b+2d+1=0$. Letting $b=0$ yields $d=-\frac{1}{2}$. Finally, we have

$$\begin{aligned} x(t) &= C_1 e^{-t} + C_2 e^{5t} - 2t e^{-t} \\ y(t) &= -C_1 e^{-t} + 2C_2 e^{5t} + (2t - \frac{1}{2}) e^{-t}. \end{aligned} \tag{2.60}$$

EXAMPLE 2.7. Find all solutions of the DE

$$3\ddot{x} + 12x = 2\sin^2 t. \tag{2.61}$$

Solution: Equation (2.61) can be rewritten as

$$\ddot{x} + 4x = \frac{1}{3}(1-\cos 2t) = \frac{1}{3} - \frac{1}{3}\cos 2t. \tag{2.62}$$

The roots of the C.P. $L(\lambda) = \lambda^2 + 4$ are $\lambda_\pm = \pm 2i$. So the basic solutions are $\cos 2t$ and $\sin 2t$. Since $f(t) = \frac{1}{3} - \frac{1}{3}\cos 2t$, so $f_1(t) = \frac{1}{3}$ and $f_2(t) = -\frac{1}{3}\cos 2t$. For $f_1(t) = \frac{1}{3}$, let $x_1^p(t) = a$, then

$$\ddot{x}_1^p + 4x_1^p = \frac{1}{3} \qquad \Longrightarrow \qquad 4a = \frac{1}{3} \qquad \Longrightarrow \qquad a = \frac{1}{12}. \tag{2.63}$$

Hence, $x_1^p(t) = \frac{1}{12}$. For $f_2(t) = -\frac{1}{3}\cos 2t$ (resonant case), let

$$x_2^p(t) = (bt + c)\cos 2t + (dt + e)\sin 2t, \tag{2.64}$$

then,

$$\begin{aligned} \dot{x}_2^p &= (2dt + 2e + b)\cos 2t - (2bt + 2c - d)\sin 2t \\ \ddot{x}_2^p &= (-4bt - 4c + 4d)\cos 2t + (-4dt - 4e - 4b)\sin 2t \end{aligned}$$

and so

$$\ddot{x}_2^p + 4x_2^p = -\frac{1}{3}\cos 2t \qquad \Longrightarrow \qquad 4d\cos 2t - 4b\sin 2t = -\frac{1}{3}\cos 2t \tag{2.65}$$

which yields

$$b = 0 \qquad \text{and} \qquad d = -\frac{1}{12}, \tag{2.66}$$

but c and e are not determined. Let $c = e = 0$ (nonzero c and e are also solutions), we have $x_2^p(t) = -\dfrac{1}{12}\, t \sin 2t$. Therefore, the general solution of (2.61) is given by

$$x(t) = C_1 \cos 2t + C_2 \sin 2t + \frac{1}{12}(1 - t \sin 2t). \tag{2.67}$$

EXAMPLE 2.8. Find all solutions of the DE

$$\ddot{x} + x = -\sin^3 t\,. \tag{2.68}$$

Solution: Similar to the above example, we rewrite (2.68) as

$$\ddot{x} + x = -\sin^3 t = -\frac{3}{4}\sin t + \frac{1}{4}\sin 3t. \tag{2.69}$$

It is clear to see that the first term of the right-hand side gives the resonant case while the second term yields non-resonant case. The roots of the C.P. $L(\lambda) = \lambda_2 + 1$ are $\lambda_\pm = \pm i$. So the basic solutions are $\cos t$ and $\sin t$. For $f_1(t) = -\dfrac{3}{4}\sin t$, we have

$$x_1^p(t) = (a\,t + b)\cos t + (c\,t + d)\sin t. \tag{2.70}$$

Then,

$$\begin{aligned} \dot{x}_1^p &= (c\,t + d + a)\cos t - (a\,t + b - c)\sin t \\ \ddot{x}_1^p &= -(a\,t + b - 2\,c)\cos t - (c\,t + d + 2\,a)\sin t \end{aligned}$$

and so

$$\ddot{x}_1^p + x_1^p = -\frac{3}{4}\sin t \qquad \Longrightarrow \qquad 2\,c\cos t - 2\,a\sin t = -\frac{3}{4}\sin t \tag{2.71}$$

which results in

$$a = \frac{3}{8} \qquad \text{and} \qquad c = 0, \tag{2.72}$$

while b and d are not determined; so let $b = d = 0$. Thus, $x_1^p(t) = \dfrac{3}{8}\, t \cos t$.

For $f_2(t) = \dfrac{1}{4}\sin 3t$, we may let

$$x_2^p(t) = e\cos 3t + h \sin 3t. \tag{2.73}$$

Then, $\ddot{x}_2^p = -9\,e\,\cos 3t - 9\ h\,\sin 3t$, so

$$\ddot{x}_2^p + x_2^p = \frac{1}{4}\,\sin 3t \qquad \Longrightarrow \qquad -8\,e\,\cos 3t - 8\,h\,\sin 3t = \frac{1}{4}\,\sin 3t \tag{2.74}$$

which yields

$$e = 0 \qquad \text{and} \qquad h = -\frac{1}{32}. \tag{2.75}$$

Thus, $x_2^p(t) = -\dfrac{1}{32}\,\sin 3t$. Hence, the general solution of the DE (2.68) is given by

$$x(t) = C_1\,\cos t + C_2\,\sin t + \frac{3}{8}\,t\,\cos t - \frac{1}{32}\,\sin 3t. \tag{2.76}$$

Consider the **Euler's equation**:

$$t^n\,\frac{d^n y}{dt^n} + a_1\,t^{n-1}\,\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_n\,y = 0 \tag{2.77}$$

where $a_i\,,\ i = 1,\,2,\,\cdots,n$ are constants. Suppose $0 < t < \infty$, let $t = e^\tau$, then

$$\begin{aligned}
\frac{dy}{dt} &= \frac{dy}{d\tau}\,\frac{d\tau}{dt} = \frac{1}{t}\,\frac{dy}{d\tau} \qquad \text{or} \qquad t\,\frac{dy}{dt} = \frac{dy}{d\tau} \\
\frac{d^2y}{dt^2} &= \frac{1}{t^2}\left(\frac{d^2y}{dt^2} - \frac{dy}{d\tau}\right) \qquad \text{or} \qquad t^2\,\frac{d^2y}{dt^2} = \frac{d^2y}{d\tau^2} - \frac{dy}{d\tau} \\
&\ \ \vdots
\end{aligned}$$

Thus, (2.77) becomes an ODE with constant coefficients, as discussed above.

EXAMPLE 2.9. Find all solutions of the DE

$$(1+t)^2\,\ddot{x} + (1+t)\,\dot{x} + x = 4\,\cos\ln(1+t). \tag{2.78}$$

Solution: First use $\bar{t} = 1 + t$ to transform (2.78) to

$$\bar{t}^2\,\frac{d^2x}{d\bar{t}^2} + \bar{t}\,\frac{dx}{d\bar{t}} + x = 4\,\cos\ln\bar{t} \tag{2.79}$$

which is in the form of Euler's equation. So let $\bar{t} = e^\tau$, we have $\bar{t}\,\dfrac{dx}{d\bar{t}} = \dfrac{dx}{d\tau}$ and $\bar{t}^2\,\dfrac{d^2x}{d\bar{t}^2} = \dfrac{d^2x}{d\tau^2} - \dfrac{dx}{d\tau}$, and thus (2.79) becomes

$$\frac{d^2x}{d\tau^2} + x = 4\,\cos\tau. \tag{2.80}$$

The basic solutions of (2.80) are $\cos\tau$ and $\sin\tau$. Here, $f(t) = 4\cos\tau$ which is the resonant case. So let

$$x^p(t) = (a\,\tau + b)\,\cos\tau + (c\,\tau + d)\,\sin\tau \tag{2.81}$$

and so

$$\begin{aligned} \frac{dx^p}{d\tau} &= (c\,\tau + d + a)\,\cos\tau - (a\,\tau + b - c)\,\sin\tau \\ \frac{d^2x^p}{d\tau^2} &= -(a\,\tau + b - 2\,c)\,\cos\tau - (c\,\tau + d + 2\,a)\,\sin\tau. \end{aligned}$$

Then,

$$\ddot{x}^p + x^p = 4\,\cos\tau \qquad \Longrightarrow \qquad 2\,c\,\cos\tau - 2\,a\,\sin\tau = 4\,\cos\tau \tag{2.82}$$

which yields

$$a = 0\,, \quad c = 2 \qquad \text{and} \qquad b = d = 0. \tag{2.83}$$

Thus, $x^p(t) = 2\,\tau\,\sin\tau$. Hence, the general solution of (2.80) is given by

$$x(\tau) = C_1\,\cos\tau + C_2\,\sin\tau + 2\,\tau\,\sin\tau. \tag{2.84}$$

Returning to the original DE (2.78), one obtains

$$x(t) = C_1\,\cos\left[\ln(1+t)\right] + C_2\,\sin\left[\ln(1+t)\right] + 2\,\ln(1+t)\,\sin\left[\ln(1+t)\right]. \tag{2.85}$$

2.4 Variation of Parameters

We will restrict to first- and second-order ODE but with non-constant coefficients. First consider the first-order linear ODE:

$$\frac{dy}{dt} + P(t)\,y = F(t). \tag{2.86}$$

When $F(t) \equiv 0$, (2.86) is a homogeneous linear ODE. The general solution of (2.86) can be found by first finding the solution of the corresponding homogeneous DE

$$\frac{dy}{dt} + P(t)\,y = 0 \tag{2.87}$$

which is *separable* and the solution is given by

$$y = k\,e^{-\int_{t_0}^{t} P(\tau)\,d\tau} \tag{2.88}$$

where the constant $k \neq 0$. Now based on the solution (2.88) we use the *method of variation of parameters* to find the solution of the DE (2.86). The idea is to suppose that the constant k is now a function $k(t)$ to be determined such that

$$y = k(t)\, e^{-\int_{t_0}^{t} P(\tau)\, d\tau} \tag{2.89}$$

is the solution of (2.86). To determine the function $k(t)$, substitute (2.89) into (2.86) to obtain

$$\frac{dk}{dt}\, e^{-\int_{t_0}^{t} P(\tau)\, d\tau} - k(t)\, P(t)\, e^{-\int_{t_0}^{t} P(\tau)\, d\tau} + P(t)\, k(t)\, e^{-\int_{t_0}^{t} P(\tau)\, d\tau} = F(t) \tag{2.90}$$

i.e.,

$$\frac{dk}{dt}\, e^{-\int_{t_0}^{t} P(\tau)\, d\tau} = F(t) \tag{2.91}$$

or

$$\frac{dk}{dt} = F(t)\, e^{\int_{t_0}^{t} P(\tau)\, d\tau}. \tag{2.92}$$

Thus,

$$k(t) = C + \int_{t_0}^{t} F(t_1)\, e^{\int_{t_0}^{t} P(\tau)\, d\tau}\, dt_1. \tag{2.93}$$

Hence,

$$y(t) = e^{-\int_{t_0}^{t} P(\tau)\, d\tau} \left(C + \int_{t_0}^{t} F(t_1)\, e^{\int_{t_0}^{t} P(\tau)\, d\tau}\, dt_1 \right) \tag{2.94}$$

is the general solution of the DE (2.86), where C is an arbitrary constant determined by the initial condition. The solution (2.94) is usually written as an indefinite integral

$$y(t) = e^{-\int P(t)\, dt} \left(C + \int F(t)\, e^{\int P(t)\, dt}\, dt \right). \tag{2.95}$$

EXAMPLE 2.10. A simple electric circuit, shown in Fig. 2.1, consists of an electric source E, a conductor L and a resistor R which are connected in series. Find the current $I(t)$ of the circuit at time t.

Solution: According to the Kirchoff Law, we have the following DE

$$E = R\, I + L\, \frac{dI}{dt}, \tag{2.96}$$

thus,

$$\frac{dI}{dt} + \frac{R}{L}\, I = \frac{E}{L}. \tag{2.97}$$

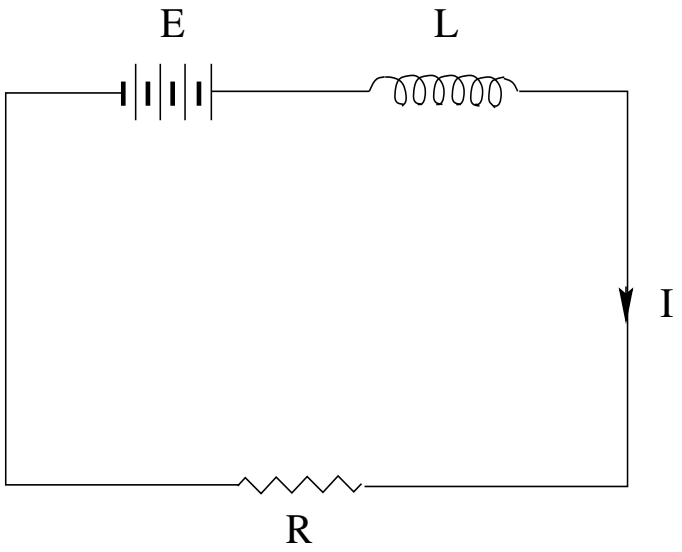


Fig. 2.1 An electric circuit.

Suppose E is a constant, I is direct current; and also assume that the current I at time t is $I(0) = I_0$. Then, the formula (2.95) gives the solution

$$I(t) = e^{-\int \frac{R}{L}\,dt}\left(C + \int \frac{E}{L}\, e^{\int \frac{R}{L}\,dt}\,dt\right) = e^{-\frac{R}{L}t}\left(C + \frac{E}{R}\, e^{\frac{R}{L}t}\right) = C\, e^{-\frac{R}{L}t} + \frac{E}{R}. \tag{2.98}$$

Substituting the initial condition $I(0) = I_0$ into (2.98) yields

$$C = I_0 - \frac{E}{R}. \tag{2.99}$$

Hence, the solution to the initial value problem (IVP) is

$$I(t) = \left(I_0 - \frac{E}{R}\right) e^{-\frac{R}{L}t} + \frac{E}{R}. \tag{2.100}$$

If $I_0 = 0$, then $I(t) = \dfrac{E}{R}\left(1 - e^{-\frac{R}{L}t}\right)$.

Next, consider the quadratic DE

$$\frac{d^2y}{dt^2} + P(t)\,\frac{dy}{dt} + Q(t)\,y = F(t) \tag{2.101}$$

with $P(t)$, $Q(t)$ and $F(t)$ continuous on some interval I. Assuming that we can find two linearly independent solutions y_1 and y_2 of the corresponding homogeneous DE of (2.101),

then the variation of parameters method attempts to find a particular solution of (2.101) of the form

$$y^p(t) = k_1(t)\, y_1(t) + k_2(t)\, y_2(t) \tag{2.102}$$

in which k_1 and k_2 are twice-differentiable functions.

We need two equations to determine the unknown functions k_1 and k_2. In fact, we will obtain equations from which we can solve for $\dot{k}_1$ and $\dot{k}_2$, and then obtain k_1 and k_2 by integration. In order to substitute $k_1(t)\, y_1(t) + k_2(t)\, y_2(t)$ into the DE (2.101), first compute

$$\frac{dy^p}{dt} = k_1(t)\, \frac{dy_1}{dt} + k_2(t)\, \frac{dy_2}{dt} + \frac{dk_1}{dt}\, y_1(t) + \frac{dk_2}{dt}\, y_2(t). \tag{2.103}$$

To simplify both $\dfrac{dy^p}{dt}$ and $\dfrac{d^2y^p}{dt^2}$, we will require that k_1 and k_2 must satisfy

$$\dot{k}_1\, y_1 + \dot{k}_2\, y_2 = 0 \tag{2.104}$$

where the dot denotes the differentiation with respect to time t, and the variable t is dropped for simplicity. This gives us one equation for $\dot{k}_1$ and $\dot{k}_2$. Now,

$$\ddot{y}^p = \dot{k}_1\, \dot{y}_1 + \dot{k}_2\, \dot{y}_2 + k_1\, \ddot{y}_1 + k_2\, \ddot{y}_2. \tag{2.105}$$

The equation $\ddot{y}^p + P\, \dot{y}^p + Q\, y^p = F$ now becomes

$$\dot{k}_1\, \dot{y}_1 + \dot{k}_2\, \dot{y}_2 + k_1\, \ddot{y}_1 + k_2\, \ddot{y}_2 + P\, (k_1\, \dot{y}_1 + k_2\, \dot{y}_2) + Q\, (k_1\, y_1 + k_2\, y_2) = F \tag{2.106}$$

which can be rewritten as

$$k_1\, (\ddot{y}_1 + P\, \dot{y}_1 + Q\, y_1) + k_2\, (\ddot{y}_2 + P\, \dot{y}_2 + Q\, y_2) + \dot{k}_1\, \dot{y}_1 + \dot{k}_2\, \dot{y}_2 = F. \tag{2.107}$$

Since y_1 and y_2 are assumed to be solutions of the associated homogeneous equation $\ddot{y} + P\, \dot{y} + Q\, y = 0$, the terms in brackets are zero, and we have

$$\dot{k}_1\, \dot{y}_1 + \dot{k}_2\, \dot{y}_2 = F, \tag{2.108}$$

a second equation for $\dot{k}_1$ and $\dot{k}_2$.

Equations (2.104) and (2.108) give us two equations for $\dot{k}_1$ and $\dot{k}_2$. For case of reference, we rewrite these two equations together

$$\begin{aligned} y_1\,\dot{k}_1 + k_2\,\dot{k}_2 &= 0 \\ \dot{y}_1\,\dot{k}_1 + \dot{y}_2\,\dot{k}_2 &= F. \end{aligned} \tag{2.109}$$

This system of equation has a unique solution if and only if the determinant of the coefficients is nonzero. This determinant is

$$W = \begin{vmatrix} y_1 & y_2 \\ \dot{y}_1 & \dot{y}_2 \end{vmatrix}, \tag{2.110}$$

where W represents the *Wronskian* W of y_1 and y_2. Since y_1 and y_2 are assumed linearly independent on I, $W(t) \neq 0$ for all $t \in I$. The solutions of (2.109) for $\dot{k}_1$ and $\dot{k}_2$ are

$$\dot{k}_1 = \frac{\begin{vmatrix} 0 & y_2 \\ F & \dot{y}_2 \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ \dot{y}_1 & \dot{y}_2 \end{vmatrix}} = -\frac{y_2\,F}{W} \qquad \text{and} \qquad \dot{k}_2 = \frac{\begin{vmatrix} y_1 & 0 \\ \dot{y}_1 & F \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ \dot{y}_1 & \dot{y}_2 \end{vmatrix}} = \frac{y_1\,F}{W}. \tag{2.111}$$

If we can carry out the necessary integrations, we have functions k_1 and k_2 such that

$$y^p = k_1\,y_1 + k_2\,y_2$$

is a particular solution of the DE $\ddot{y} + P\,\dot{y} + Q\,y = F$. Usually we let the constants of integration be zero in integrating to obtain k_1 and k_2 because we do not need more than one k_1 and k_2 having the required properties.

EXAMPLE 2.11. Find a solution of the DE

$$\ddot{y} + 4\,y = \tan(2\,t) \tag{2.112}$$

on the interval $\left(-\dfrac{\pi}{4},\ \dfrac{\pi}{4}\right)$.

Solution: The associated homogeneous equation $\ddot{y} + 4\,y = 0$ has $y_1(t) = \cos(2\,t)$ and $y_2(t) = \sin(2\,t)$ as a fundamental set of solutions. Their Wronskian is

$$W(t) = \begin{vmatrix} \cos(2\,t) & \sin(2\,t) \\ -2\,\sin(2\,t) & 2\,\cos(2\,t) \end{vmatrix} = 2\,\cos^2(2\,t) + 2\,\sin^2(2\,t) = 2. \tag{2.113}$$

With $F(t) = \tan(2\,t)$, equation (2.111) gives us

$$\begin{aligned} \dot{k}_1 &= -\frac{y_2(t)\,F(t)}{W(t)} = -\frac{1}{2}\,\sin(2\,t)\,\tan(2\,t) \\ \dot{k}_2 &= \frac{y_1(t)\,F(t)}{W(t)} = \frac{1}{2}\,\cos(2\,t)\,\tan(2\,t) \end{aligned} \tag{2.114}$$

which yields

$$\begin{aligned} k_1(t) &= \int -\frac{1}{2}\,\sin(2\,t)\,\tan(2\,t)\,dt = -\frac{1}{2}\int \frac{\sin^2(2\,t)}{\cos(2\,t)}\,dt \\ &= -\frac{1}{2}\int \frac{1-\cos^2(2\,t)}{\cos(2\,t)}\,dt = -\frac{1}{2}\int \sec(2\,t)\,dt + \frac{1}{2}\int \cos(2\,t)\,dt \\ &= -\frac{1}{4}\,\ln|\sec(2\,t)+\tan(2\,t)| + \frac{1}{4}\,\sin(2\,t) \end{aligned} \tag{2.115}$$

and

$$k_2(t) = \frac{1}{2}\int \cos(2\,t)\,\tan(2\,t)\,dt = \frac{1}{2}\int \sin(2\,t)\,dt = -\frac{1}{4}\,\cos(2\,t). \tag{2.116}$$

A particular solution of (2.112) is

$$\begin{aligned} y^p(t) &= k_1(t)\,y_1(t) + k_2(t)\,y_2(t) \\ &= \Big[-\frac{1}{4}\,\ln|\sec(2\,t)+\tan(2\,t)| + \frac{1}{4}\,\sin(2\,t)\Big]\cos(2\,t) - \frac{1}{4}\,\cos(2\,t)\,\sin(2\,t) \\ &= -\frac{1}{4}\,\ln|\sec(2\,t)+\tan(2\,t)|\,\cos(2\,t). \end{aligned} \tag{2.117}$$

Hence, the general solution of (2.112) on the interval $\left(-\frac{\pi}{4},\ \frac{\pi}{4}\right)$ is given by

$$y(t) = C_1\,\cos(2\,t) + C_2\,\sin(2\,t) - \frac{1}{4}\,\ln|\sec(2\,t)+\tan(2\,t)|\,\cos(2\,t). \tag{2.118}$$

Although the DE in Example 2. 11 has constant coefficients, this is not a requirement for the method of variation of parameters, as it was for the method of undetermined coefficients.

EXAMPLE 2.12. Find a solution of the DE

$$\ddot{y} - \frac{4}{t}\,\dot{y} + \frac{4}{t^2}\,y = t^2 + 1 \qquad \text{for } \ t > 0. \tag{2.119}$$

Solution: The associated homogeneous equation $\ddot{y} - \frac{4}{t}\,\dot{y} + \frac{4}{t^2}\,y = 0$ is a Euler equation, and we find the fundamental set of solutions $y_1(t) = t$ and $y_2(t) = t^4$ for $t > 0$. The Wronskian of y_1 and y_2 is

$$W(t) = \begin{vmatrix} t & t^4 \\ 1 & 4\,t^3 \end{vmatrix} = 4\,t^4 - t^4 = 3\,t^4 \tag{2.120}$$

and this is nonzero for $t > 0$. With $F(t) = t^2 + 1$, equation (2.111) gives us

$$\begin{aligned} \dot{k}_1 &= -\frac{t^4\,(t^2+1)}{3\,t^4} = -\frac{t^2+1}{3} \\ \dot{k}_2 &= \frac{t\,(t^2+1)}{3\,t^4} = \frac{1}{3\,t} + \frac{1}{3\,t^3} \end{aligned} \tag{2.121}$$

which yields

$$\begin{aligned} k_1(t) &= \int -\frac{t^2+1}{3}\,dt = -\frac{1}{9}\,t^3 - \frac{1}{3}\,t \\ k_2(t) &= \int \left[\frac{1}{3\,t} + \frac{1}{3\,t^3}\right] dt = \frac{1}{3}\,\ln(t) - \frac{1}{6\,t^2}. \end{aligned} \tag{2.122}$$

A particular solution of (2.119) is

$$\begin{aligned} y^p(t) &= \left(-\frac{1}{9}\,t^3 - \frac{1}{3}\,t\right) t + \left(\frac{1}{3}\,\ln(t) - \frac{1}{6\,t^2}\right) t^4 \\ &= -\frac{1}{9}\,t^4 - \frac{1}{2}\,t^2 + \frac{1}{3}\,t^4\,\ln(t). \end{aligned} \tag{2.123}$$

Hence, the general solution of (2.119) is

$$y(t) = C_1\,t + C_2\,t^4 - \frac{1}{9}\,t^4 - \frac{1}{2}\,t^2 + \frac{1}{3}\,t^4\,\ln(t). \tag{2.124}$$

We have considered the method of variation of parameters for the DE

$$\ddot{y} + P(t)\,\dot{y} + Q(t)\,y = F(t).$$

Although P and Q are not be constant to use the method, we have assumed that the coefficients of $\ddot{y}$ is 1. If we encounter a DE in which the coefficient of $\ddot{y}$ is a nonconstant function $R(t)$, we must divide the equation by $R(t)$ (in an interval in which $R(t) \neq 0$) before we can apply the method to the resulting equation.

Chapter 3

SOME "SOLVABLE" TYPES OF NONLINEAR ODE

In this chapter we will discuss some "solvable" nonlinear ordinary differential equations (NLODE). By "solvable" we mean that one can use some techniques to find the analytical solutions of the NLODE.

3.1 Classification

Before we can begin, in practice we must identify what makes a DE linear or nonlinear. A DE (ordinary as opposed to ones involving partial derivatives) of *order (n)* can generally be written in the form

$$F\left(x,\, y,\, y',\, y'',\, \cdots,\, y^{(n)}\right) = 0 \qquad \text{where} \quad y^{(k)} = \frac{d^k y}{dx^k}\,,\ k = 1,\, 2,\, \cdots,\, n \tag{3.1}$$

which is a function of x, y and all its derivatives. For example,

$$F\left(x,\, y,\, y',\, y''\right) = y'' + 2\,(y')^2 + \sin x^3 = 0 \tag{3.2}$$

where x is *independent variable* and y is *dependent variable.*

If F is nonlinear in the dependent variable, y, or its derivatives, then the DE (3.1) is said to be nonlinear. However, nonlinearity in the independent variable, x, is irrelevant. Thus,

$$y' + x^2\, y = 0 \tag{3.3}$$

is linear – but not in x.

Sometimes for first-order equations (considering one may choose which is independent and dependent for first-order equations), one can convert a seemingly nonlinear equation into a linear equation in the other variable. For example,

$$y' + \frac{y^2}{x} = 0 \tag{3.4}$$

is a *nonlinear* DE if x is chosen as independent variable while y is chosen as dependent variable. However, if we choose y as independent variable while x as dependent variable, then the DE (3.4) is *linear* since we can rewrite (3.4) in the form

$$\frac{dx}{dy} + \frac{x}{y^2} = 0 \tag{3.5}$$

which is a *linear* ODE. Such possibilities are not normally available, of course, for higher order equations.

The **degree** of F is the degree (algebraic) of $y^{(n)}$ in F. Thus,

$$y'' + 2\,(y')^2 + y + \sin x^3 = 0 \tag{3.6}$$

is a *first degree, second order* differential equation; while

$$(y')^2 + 2\,x^2\,y + x = 0 \tag{3.7}$$

is a *first order, second degree* differential equation.

Other nonlinear DEs may, of course, involve *transcendental* functions of y, for example, the simple (mathematical) *pendulum* described by

$$y'' + \sin y = 0. \tag{3.8}$$

Now let's consider an example of "solvable" types of nonlinear equations. Suppose the equation is given by

$$F(x,\, y,\, y') = 0. \tag{3.9}$$

If one can treat y' as a variable to be solved algebraically one can often convert the equation to a simple 1st-order ODE.

EXAMPLE 3.1. Solve the DE

$$F(x,\, y,\, y') = x\,y\left(\frac{dy}{dx}\right)^2 + (x+y)\,\frac{dy}{dx} + 1 = 0. \tag{3.10}$$

Solution: We can factor (3.10) in terms of $\dfrac{dy}{dx}$ as

$$\left(x\,\frac{dy}{dx} + 1\right)\left(y\,\frac{dy}{dx} + 1\right) = 0 \tag{3.11}$$

from which we have *two* equations actually

$$\left(x\,\frac{dy}{dx}+1\right)=0 \quad \text{or} \quad \left(y\,\frac{dy}{dx}+1\right)=0, \tag{3.12}$$

both of them are linear ODEs. From the 1st one, we have $\dfrac{dy}{dx}=-\dfrac{1}{x}$ and its solution is

$$\begin{aligned} y &= -\ln|x|+C_1 \\ &= -\ln e^{-C_1}\,|x| \\ &= -\ln|C'x| \qquad (C'=e^{-C_1}>0). \end{aligned} \tag{3.13}$$

From the 2nd equation, one obtains

$$y\,dy=-\,dx \qquad \Longrightarrow \qquad y^2=-\,2\,x+C_2\,. \tag{3.14}$$

Thus, we have actually *two* distinct families of solutions. In other words, for each point in the $(x,\,y)$ plane, there exist two different solution curves passing through this point. See Fig. 3.1 for the case when $C_1=2$ and $C_2=6$.

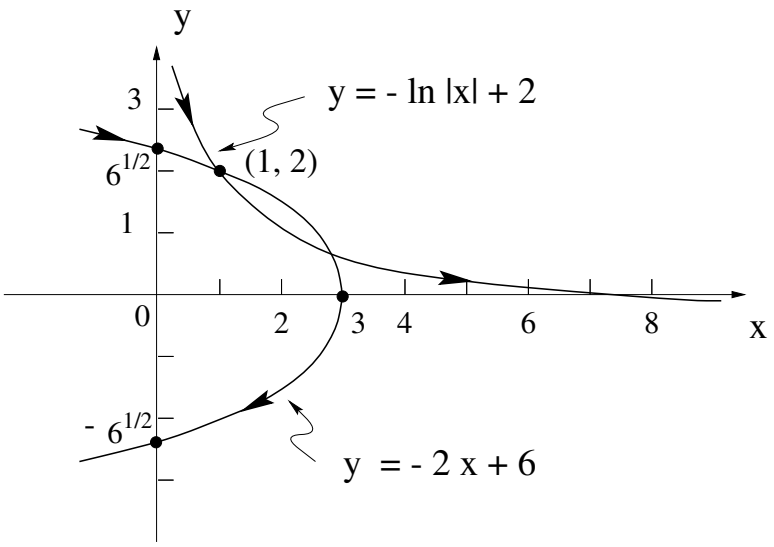


Fig. 3.1 Two solution curves of DE (3.10).

Note that at the intersection point $(1,\,2)$ the slopes, $\dfrac{dy}{dx}$, of the two curves are different. If a solution is transited from one type to another type, the transition must occur at the

points where $\dfrac{dy}{dx}$ of these two solutions are same, for this example, that is,

$$-\frac{1}{x} = -\frac{1}{y} \qquad \text{or} \qquad y = x \tag{3.15}$$

a straight line. For example, Fig. 3.2 shows the case that the two solution curves are tangent at the point $(1,\ 1)$ at which they have the same slope -1.

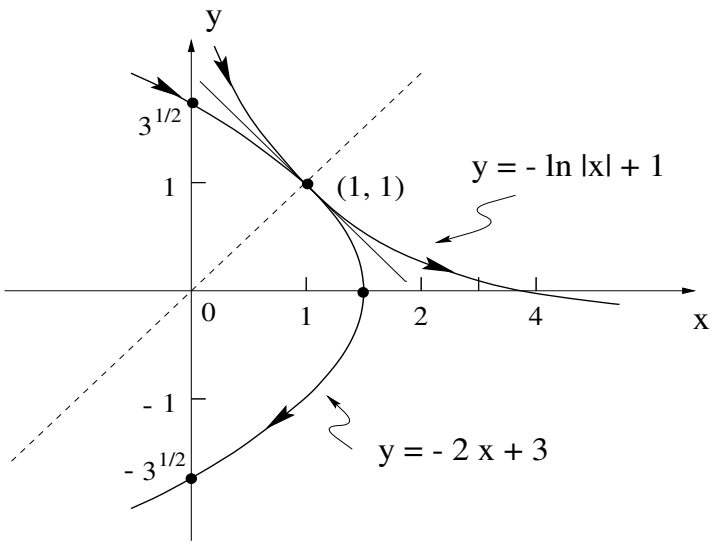


Fig. 3.2 Two tangent solution curves of DE (3.10).

3.2 Singular Solutions

For some DEs there exist some particular solution which does not belong to family of solutions of the DEs. This solution is called *singular solution*. Let's look an example. Consider

$$y^2\left(\frac{dy}{dx}\right)^2 - a^2 + y^2 = 0 \qquad (a \text{ is a constant}). \tag{3.16}$$

This DE can be rewritten as

$$\left(y\,\frac{dy}{dx}\right)^2 = a^2 - y^2 \tag{3.17}$$

and so

$$\frac{y\,dy}{\sqrt{a^2 - y^2}} = dx \tag{3.18}$$

$$\frac{y\,dy}{\sqrt{a^2 - y^2}} = -\,dx \tag{3.19}$$

or (this should not be ignored)

$$a^2 - y^2 = 0 \tag{3.20}$$

which is an important equation. By applying direct antidifferentiation to (3.18), we have

$$C_1 - \sqrt{a^2 - y^2} = x \qquad (C_1 \text{ is an arbitrary constant}) \tag{3.21}$$

which represents left-hand semi-circles. Similarly, from (3.19) one obtains

$$C_2 + \sqrt{a^2 - y^2} = x \qquad (C_2 \text{ is an arbitrary constant}) \tag{3.22}$$

which denotes right-hand semi-circles. Equations (3.21) and (3.22) can be written as

$$(x - C_1)^2 + y^2 = a^2 \qquad \text{and} \qquad (x - C_2)^2 + y^2 = a^2,$$

respectively, and these two equations can be combined to single equation

$$(x - C)^2 + y^2 = a^2. \tag{3.23}$$

Another solution of (3.16) is solved from (3.20) as

$$y = \pm a \tag{3.24}$$

which can be verified by substituting it into (3.16) to show that

$$\text{LHS} = a^2\,(0) - a^2 + a^2 = 0 = \text{RHS}.$$

Hence, the *general solution* (3.23) represents circles centered on the x-axis (depending upon the choice of C. See Fig. 3.3.) It is seen that any point on solution $y = a$ or $y = -a$ is on a particular circle and the particular curve (here, line) is tangent to the circle at this point. However, the general solution, the family of circles (or family of semi-circles) does not include the particular solution $y = \pm a$. From the viewpoint of geometry, this particular curve $y = a$ or $y = -a$ is the envelope of the family of circles. While from the viewpoint of differential equations, at each point of the particular solution, the *uniqueness of solution* is violated. (We shall discuss more about uniqueness of solution later.) These ($y = \pm a$) are special class of solutions known as *singular solutions* that only appear in nonlinear DEs.

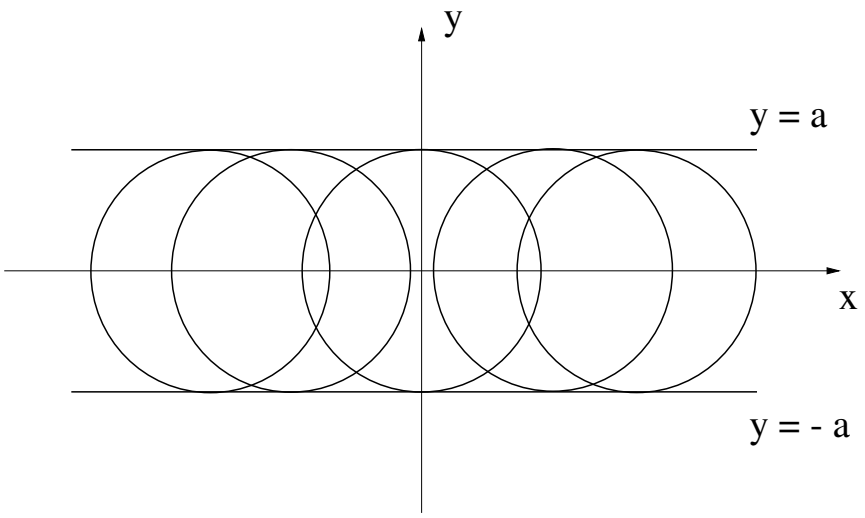


Fig. 3.3 Solutions of the DE (3.16).

Definition 3.1: A *singular solution* is the one corresponding to an integral curve that violates uniqueness of solutions at any point on the curve.

In first-order DEs these solutions must be tangent to some element of the solutions from the one-parameter family.

Next, let's discuss how to find singular solutions for general first-order nonlinear differential equations.

3.3 C-Discriminant Equation

Suppose that a general nonlinear ODE is given by

$$F(x,\, y,\, \frac{dy}{dx}) = 0 \tag{3.25}$$

where F is not necessarily a polynomial in x, y and $\dfrac{dy}{dx}$; and if it is a polynomial, it is not necessary possible to factor it. We further assume that a solution of (3.25) exists (we'll discuss existence of solutions later) which we can write in the implicit form

$$\phi(x,\, y,\, C) = 0 \tag{3.26}$$

where C is an arbitrary constant. If a singular solution exists, it must be on the envelope of the solutions as in the case of the example of circles discussed in the previous section. That

is because a singular solution must be tangent to the solutions somewhere. Each point of the envelope must be a curve dependent parametrically only on C; i.e.,

$$x = x(C) \ , \quad y = y(C). \tag{3.27}$$

But x and y are still on the curves of family of solutions, so they must satisfy

$$\phi(x(C),\, y(C), C) = 0. \tag{3.28}$$

Thus, in general

$$d\phi = 0 \qquad \Longrightarrow \qquad \frac{\partial \phi}{\partial x}\, dx + \frac{\partial \phi}{\partial y}\, dy + \frac{\partial \phi}{\partial C}\, dC = 0 \tag{3.29}$$

must hold. But on any member of the family of solutions, $dC = 0$. Hence,

$$\frac{\partial \phi}{\partial x}\, dx + \frac{\partial \phi}{\partial y}\, dy = 0. \tag{3.30}$$

This must be true even at points where a family member touches the envelope. Therefore, on the envelope where $dC \neq 0$, we must have $\dfrac{\partial \phi}{\partial C} = 0$ to satisfy $d\phi = 0$. Hence, we have two equations:

$$\phi = 0 \qquad \text{and} \qquad \frac{\partial \phi}{\partial C} = 0 \tag{3.31}$$

which must be satisfied on the envelope. From these two equations, we can get so called *C-discriminant equation* by eliminating the parameter C. This equation might represent a singular solution. That is, this C-discriminant equation needs not necessarily be a singular solution. In other words, these two equations are necessary conditions, but not sufficient! It means that **if the family of solutions has any envelope, they must be included in the C-discriminant equation, while a C-discriminant equation might not represent an envelope**.

Now let's use C-discriminant equation to consider the example given by equation (3.16).

EXAMPLE 3.2. Find possible singular solutions of the DE

$$y^2 \left(\frac{dy}{dx}\right)^2 - a^2 + y^2 = 0. \tag{3.32}$$

Solution: From the discussion given in the previous section we know that the general solution of the DE is

$$\phi(x,\, y,\, C) = (x - C)^2 + y^2 - a^2 = 0. \tag{3.33}$$

So, using

$$\frac{\partial \phi}{\partial C} = -\,2\,(x - C) = 0 \tag{3.34}$$

to eliminate C from (3.33) yields

$$y^2 - a^2 = 0 \qquad \Longrightarrow \qquad y = \pm\, a. \tag{3.35}$$

We have seen that $y = a$ or $y = -\,a$ is indeed the envelope of the family of solutions, i.e., circles

$$(x - C)^2 + y^2 - a^2 = 0.$$

Hence, $y = \pm\, a$ are singular solutions. Consider another example.

EXAMPLE 3.3. Find possible singular solutions of the DE

$$y^{\frac{2}{3}} \left(\frac{dy}{dx}\right) - \frac{3}{5} = 0. \tag{3.36}$$

Solution: It is easy to find the solution by direct antidifferentiation as

$$\phi(x,\, y,\, C) = y^5 - (x + C)^3 = 0 \qquad \text{i.e.} \quad y = (x + C)^{\frac{3}{5}}. \tag{3.37}$$

So, we have another equation

$$\frac{\partial \phi}{\partial C} = 3\,(x + C)^2 = 0 \tag{3.38}$$

Eliminating C from the above two equations results in the C-discriminant equation

$$y = 0. \tag{3.39}$$

However, $y = 0$ is not the solution of the DE (3.36), therefore, it is not a singular solution of (3.36). Indeed, Fig. 3.4. shows that the family of solutions has no envelopes.

Other family of solutions which do not have envelopes are circles having a same center and parallel lines (Fig. 3.5).

Note: Something called the *p-discriminant equation* arises in the case of repeated roots of the DE in $\dfrac{dy}{dx}$.

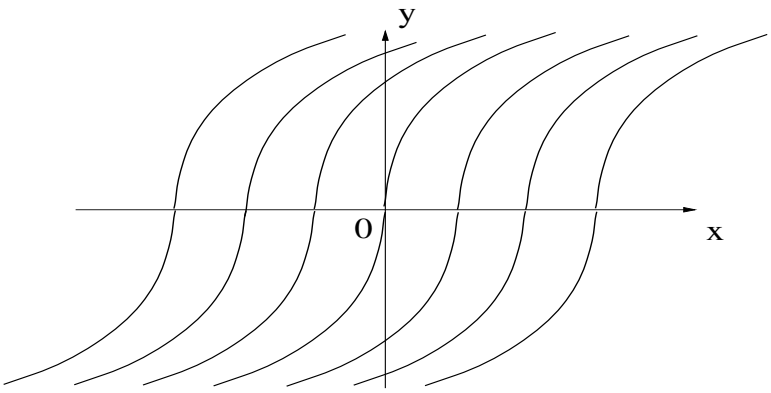


Fig. 3.4 Solutions for Example 3.3.

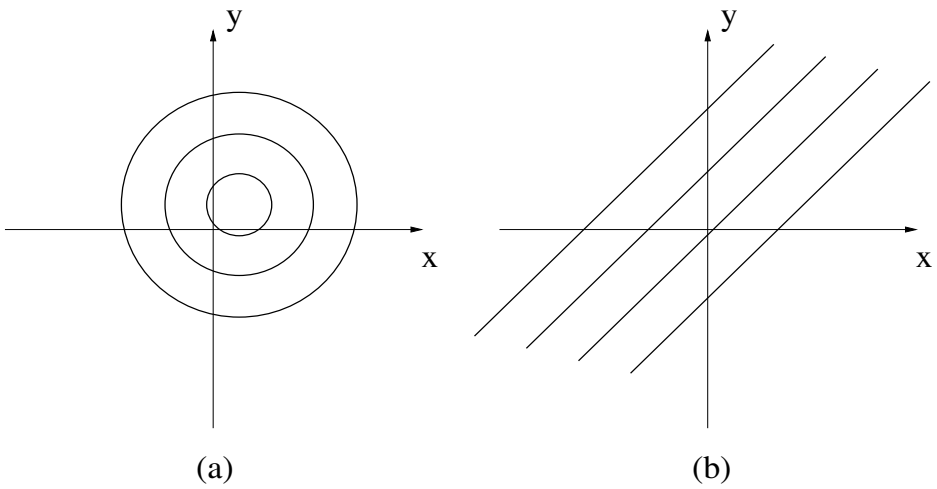


Fig. 3.5 Examples having no envelopes.

3.4 Eliminating Dependent Variable

Again consider the DE

$$F(x,\, y,\, \frac{dy}{dx}) = 0.$$

Suppose we could write this equation in the form (of course, maybe we can't!):

$$y = g(x,\, \frac{dy}{dx}). \tag{3.40}$$

Letting $p = \dfrac{dy}{dx}$ transforms (3.40) to

$$y = g(x,\, p), \tag{3.41}$$

and then differentiating this equation with respect to x yields

$$\frac{dy}{dx} = \frac{\partial g}{\partial x} + \frac{\partial g}{\partial p}\frac{dp}{dx} = h(x,\, p,\, \frac{dp}{dx}) \tag{3.42}$$

in which the dependent variable y has been eliminated. Thus, we have

$$p = h(x,\, p,\, \frac{dp}{dx}). \tag{3.43}$$

Solving for p from (3.43) may be easier than solving for y from the original DE (3.40).

EXAMPLE 3.4. Solve the DE

$$x\left(\frac{dy}{dx}\right)^2 - 3\,y\left(\frac{dy}{dx}\right) + 9\,x^2 = 0 \qquad \text{for} \quad x > 0. \tag{3.44}$$

Solution: Let $p = \dfrac{dy}{dx}$, then (3.44) is transformed to

$$x\,p^2 - 3\,y\,p + 9\,x^2 = 0 \tag{3.45}$$

from which we can separate y out to obtain

$$3\,y = x\,p + \frac{9\,x^2}{p}. \tag{3.46}$$

Next, differentiating (3.46) with respect to x yields

$$3\frac{dy}{dx} = p + x\frac{dp}{dx} + \frac{18\,x}{p} - \frac{9\,x^2}{p^2}\frac{dp}{dx}, \tag{3.47}$$

which can be written, by using $\dfrac{dy}{dx} = p$, as

$$3\,p = p + \frac{18\,x}{p} + \left(x - \frac{9\,x^2}{p^2}\right)\frac{dp}{dx}, \tag{3.48}$$

or

$$2\,p\left(1 - \frac{9\,x}{p^2}\right) = x\left(1 - \frac{9\,x}{p^2}\right)\frac{dp}{dx}. \tag{3.49}$$

So, from (3.49) we have

$$\left(2\,p - x\,\frac{dp}{dx}\right)\left(1 - \frac{9\,x}{p^2}\right) = 0 \tag{3.50}$$

which yields either

$$2\,p - x\,\frac{dp}{dx} = 0 \tag{3.51}$$

or

$$1 - \frac{9\,x}{p^2} = 0. \tag{3.52}$$

Equation (3.51) is separable and its solution can be obtained as follows:

$$\frac{dp}{p} = 2\,\frac{dx}{x}$$

$$\Longrightarrow$$

$$\ln|p| = 2\,\ln|x| + C_1 = \ln x^2 + C_1$$

$$\Longrightarrow$$

$$|p| = e^{C_1}\,x^2 = C_2\,x^2 \qquad (C_2 = e^{C_1} > 0)$$

$$\Longrightarrow$$

$$p = \pm\,C_2\,x^2 = C_3\,x^2 \qquad (C_3 = \pm\,C_2 \neq 0). \tag{3.53}$$

Substituting (3.53) into (3.46) gives us

$$y = \frac{C_3}{3}\,x^3 + \frac{3}{C_3} = C\,x^3 + \frac{1}{C} \qquad \left(C = \frac{C_3}{3} \neq 0\right) \tag{3.54}$$

which is a one-parameter family of solutions.

Next, solving (3.52) yields

$$p = \pm\,3\,x^{\frac{1}{2}} \tag{3.55}$$

and then substituting it into (3.46) produces

$$3\,y = x\left(\pm\,3\,x^{1/2}\right) + \frac{9\,x^2}{(\pm\,3\,x^{1/2})} \qquad \Longrightarrow \qquad y = \pm\,2\,x^{\frac{3}{2}}. \tag{3.56}$$

Now we want to show that (3.56) is actually the C-discriminant equation of (3.54). Thus, consider the following two equations

$$\begin{aligned} \phi(x,\,y,\,C) &= y - C\,x^3 - \frac{1}{C} = 0 \\ \frac{\partial\phi}{\partial C} &= -\,x^3 + \frac{1}{C^2} = 0. \end{aligned} \tag{3.57}$$

The second equation gives us $C = \pm\,x^{-3/2}$. Substitute it into the first equation to obtain

$$y = \pm\,x^3\,x^{-3/2} \pm \frac{1}{x^{-3/2}} = \pm\,2\,x^{\frac{3}{2}}. \tag{3.58}$$

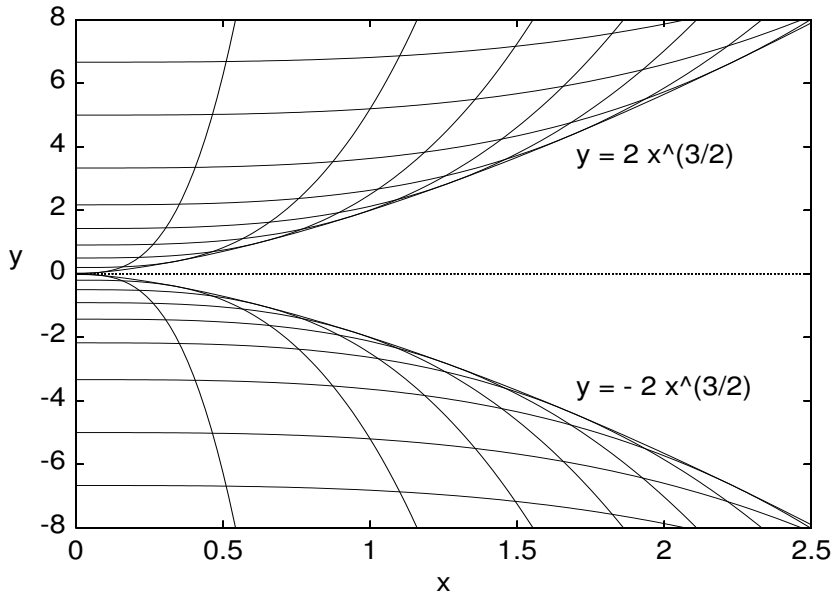


Fig. 3.6 Solutions for Example 3.4.

Fig. 3.6 shows that the C-discriminant equation is actually the envelope of the family of solutions (3.54). Hence, $y = \pm 2\,x^{3/2}$ are singular solutions of the original DE (3.44). A simple way to show that a C-discriminant equation is a singular solution is as follows: Check if the C-discriminant equation satisfies the original DE; if it does, then it is a singular solution, otherwise it is not. For Example 3.4, we know from (3.56) that $y = \pm 2\,x^{3/2}$ are solutions of (3.44), and indeed substituting $y = \pm 2\,x^{3/2}$ into (3.44) results in

$$\text{LHS} = x\left(\pm 3\,x^{\frac{1}{2}}\right)^2 - 3\left(\pm 2\,x^{\frac{3}{2}}\right)\left(\pm 3\,x^{\frac{1}{2}}\right) + 9\,x^2 = 9\,x^2 - 18\,x^2 + 9\,x^2 = 0 = \text{RHS},$$

so, $y = \pm 2\,x^{3/2}$ are singular solution of the original DE (3.44).

3.5 Eliminating Independent Variable

Similarly, the procedure discussed in the previous section can be applied to independent variable. Suppose we can write the DE $F(x,\, y,\, \dfrac{dy}{dx}) = 0$ in the form

$$x = g(y,\, p) \tag{3.59}$$

where $p = \dfrac{dy}{dx}$. Then differentiating (3.59) with respect to y with the aid $\dfrac{dx}{dy} = \dfrac{1}{p}$ results

in

$$\frac{1}{p} = \frac{\partial g}{\partial y} + \frac{\partial g}{\partial p}\,\frac{dp}{dy} = h(y,\, p,\, \frac{dp}{dy}) \tag{3.60}$$

which is a first-order DE with y as independent variable and p as dependent variable, and derivative $\dfrac{dp}{dy}$ has been solved out. Then solving (3.59) and (3.60) together yields the solution for the original DE.

EXAMPLE 3.5. Solve the DE

$$\left(\frac{dy}{dx}\right)^3 - 4\,x\,y\left(\frac{dy}{dx}\right) + 8\,y^2 = 0. \tag{3.61}$$

Solution: Letting $p = \dfrac{dy}{dx}$ and solving x from (3.61) give us

$$x = \frac{p^2}{4\,y} + \frac{2\,y}{p}. \tag{3.62}$$

Then differentiating (3.62) with respect to y yields

$$\frac{dx}{dy} = \frac{1}{p} = \left(\frac{p}{2\,y} - \frac{2\,y}{p^2}\right)\frac{dp}{dy} - \left(\frac{p^3}{4\,y^2} - \frac{2}{p}\right), \tag{3.63}$$

or

$$\frac{p^3 - 4\,y^4}{2\,y\,p^2}\left(\frac{dp}{dy} - \frac{p}{2\,y}\right) = 0. \tag{3.64}$$

From the second factor of (3.64) $\dfrac{dp}{dy} - \dfrac{p}{2\,y} = 0$ we find

$$p = c\;|y|^{\frac{1}{2}}\,. \tag{3.65}$$

Substitute (3.65) (remembering $p = \dfrac{dy}{dx}$) into the original DE (3.61) to obtain

$$c^3\,|y|^{\frac{3}{2}} - 4\,c\,x\,y\;|y|^{\frac{1}{2}} + 8\,y^2 = 0 \tag{3.66}$$

which is a one-parameter family of solutions of the DE (3.61). When $y > 0$, canceling factor $y^{\frac{3}{2}}$ in (3.66) yields $c^3 - 4\,c\,x + 8\,y^{\frac{1}{2}} = 0$; i.e.,

$$64\,y = c^2\left(4\,x - c^2\right)^2$$

or

$$y = C_1\,(x - C_1)^2 \qquad (C_1 = \frac{c^2}{4}). \tag{3.67}$$

When $y < 0$, similarly canceling factor $(-y)^{\frac{3}{2}}$ in (3.66) leads to $c^3 + 4\,c\,x + 8\,(-y)^{\frac{1}{2}} = 0$; i.e.,

$$64\,y = -\,c^2\left(4\,x + c^2\right)^2$$

or

$$y = C_2\,(x - C_2)^2 \qquad (C_2 = -\,\frac{c^2}{4}). \tag{3.68}$$

It can be seen that (3.68) actually represents (3.67) if C_1 takes negative values. Moreover, $y = 0$ is obviously a solution of (3.61). Hence

$$y = C\,(x - C)^2 \tag{3.69}$$

is family of solutions of (3.61), where C is an arbitrary constant.

Next, from the first factor of (3.64) $p^3 - 4\,y^2 = 0$ one obtains

$$p = \left(4\,y^2\right)^{\frac{1}{3}}. \tag{3.70}$$

Substituting (3.70) into (3.61) results in

$$y\,\left[3\,y - x\,\left(4\,y^2\right)^{1/3}\right] = 0.$$

So we have

$$y = 0 \qquad \text{or} \qquad y = \frac{4}{27}\,x^3. \tag{3.71}$$

On the other hand, from (3.69) we have

$$\begin{aligned} \phi &= y - C\,(x - C)^2 = 0 \\ \frac{\partial \phi}{\partial C} &= (x - C)\,(3\,C - x) = 0 \end{aligned} \tag{3.72}$$

from which eliminating C gives us

$$y = 0 \qquad \text{or} \qquad y = \frac{4}{27}\,x^3.$$

Hence, both $y = 0$ and $y = \frac{4}{27}\,x^3$ are singular solutions of the original DE (3.61).

3.6 First-Order Equations: General Strategies

In this section, we shall introduce several methods which can be used to solve for first-order DEs analytically.

3.6.1 Separate equations. Suppose algebraic manipulations of the DE $F(x,\, y,\, \dfrac{dy}{dx}) = 0$ leads to the form

$$f(x)\,dx + g(y)\,dy = 0. \tag{3.73}$$

Then its solution can be obtained by just direct integration as

$$\int f(x)\,dx + \int g(y)\,dy = C. \tag{3.74}$$

EXAMPLE 3.6. Solve the DE

$$(1+y^2)\,dx + (1+x^2)\,dy = 0. \tag{3.75}$$

Solution: (3.75) can be written as

$$\frac{dx}{1+x^2} + \frac{dy}{1+y^2} = 0. \tag{3.76}$$

So, its solution is

$$\tan^{-1} x + \tan^{-1} y = C' \tag{3.77}$$

from which we have

$$\tan\left(\tan^{-1} x + \tan^{-1} y\right) = \tan C' = C$$

$$\Longrightarrow$$

$$x + y = C\,(1 - x\,y). \tag{3.78}$$

EXAMPLE 3.7. Find all the solutions of the DE

$$\frac{dy}{dx} + \frac{x}{y} = 0. \tag{3.79}$$

Solution: We need to discuss this equation in the two regions:

$$R_1: \quad -\infty < x < \infty, \quad 0 < y < +\infty;$$

$$R_2: \quad -\infty < x < \infty, \quad -\infty < y < 0,$$

separately. In region R_1, from

$$\int y\,dy = -\int x\,dx + \frac{C^2}{2}, \tag{3.80}$$

we have the solution

$$y^2 = C^2 - x^2 \qquad \text{or} \qquad y = \sqrt{C^2 - x^2}\,, \;\; -C < x < C \tag{3.81}$$

where C is an arbitrary positive constant. Similarly,

$$y = -\sqrt{C^2 - x^2}\,, \;\; -C < x < C \tag{3.82}$$

is the solution of (3.79) in the region R_2.

EXAMPLE 3.8. Find all the solutions of the DE

$$\frac{dy}{dx} = \frac{2}{x^2 - 1}. \tag{3.83}$$

Solution: We need to discuss this equation in the three regions:

$$\begin{aligned}
&\Gamma_1: \quad -\infty < x < -1, \quad -\infty < y < +\infty;\\
&\Gamma_2: \quad -1 < x < +1, \quad -\infty < y < +\infty;\\
&\Gamma_3: \quad 1 < x < +\infty, \quad -\infty < y < +\infty,
\end{aligned}$$

separately. For each point $(x,\, y)$ of the three regions, there exists one and only one solution curve passing the point. Equation (3.83) can be written as

$$\frac{dy}{dx} = \frac{1}{x-1} - \frac{1}{x+1}. \tag{3.84}$$

Integrate the above equation to obtain

$$y = \int \frac{dx}{x-1} - \int \frac{dx}{x+1} + C = \ln\left|\frac{x-1}{x+1}\right| + C. \tag{3.85}$$

Thus,

$$\begin{aligned}
y &= \ln\frac{x-1}{x+1} + C \qquad \text{for} \qquad -\infty < x < -1 \quad \text{and} \quad 1 < x < +\infty;\\
y &= \ln\frac{1-x}{1+x} + C \qquad \text{for} \qquad -1 < x < -1
\end{aligned} \tag{3.86}$$

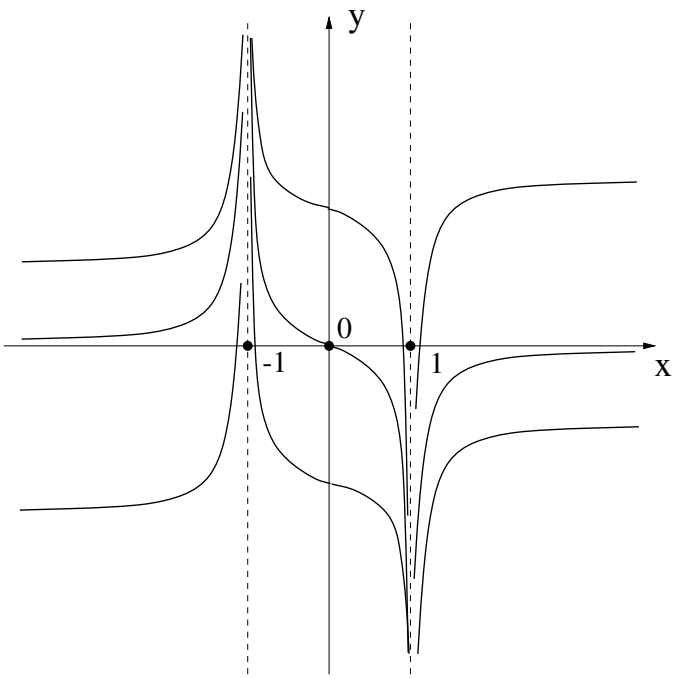


Fig. 3.7 Solutions for Example 3.8.

which implies that for a given constant C, there are three solution curves defined by (3.86) (Fig. 3.8).

EXAMPLE 3.9. Find all the solutions of the DE

$$\frac{dy}{dx} = \frac{y^2 - 1}{2}. \tag{3.87}$$

Solution: We need to discuss this equation in the three regions: $-\infty < y < -1$, $-1 < y < 1$ and $1 < x < +\infty$. Integrating (3.87) yields

$$\ln \left| \frac{y-1}{y+1} \right| = x + C'. \tag{3.88}$$

In region $-\infty < y < -1$ and $1 < y < +\infty$, (3.88) is

$$\frac{y-1}{y+1} = e^{C'} e^x = C_1 e^x \qquad (C_1 = e^{C'} > 0). \tag{3.89}$$

So the solution is

$$y = \frac{1 + C_1 e^x}{1 - C_1 e^x}. \tag{3.90}$$

When $-\infty < x < -\ln C_1,\ y > 1$; when $-\ln C_1 < x < +\infty,\ y < -1$. Note that none of the domains of these two solutions is $-\infty < x < +\infty$.

In the region $-1 < y < 1$, the solution of (3.87) is

$$\frac{y-1}{y+1} = C_2\, e^x \qquad (C_2 < 0) \tag{3.91}$$

which, in turn, gives

$$y = \frac{1 + C_2\, e^x}{1 - C_2\, e^x} \qquad (C_2 < 0). \tag{3.92}$$

Note that (3.92) is defined in $-\infty < x < +\infty$. The two solutions (3.90) and (3.92) can be combined as

$$y = \frac{1 + C\, e^x}{1 - C\, e^x} \qquad (C \neq 0). \tag{3.93}$$

But $y = -1$ and $y = 1$ are obviously also the solutions of (3.87), the former of which can be obtained from (3.93) by letting $C \to \infty$ while the latter can be got by setting $C = 0$ in (3.93). Therefore, letting the constant C varying from $-\infty$ to $+\infty$, one obtains all the solutions of (3.87) from equation (3.93). See Fig. 3.8.

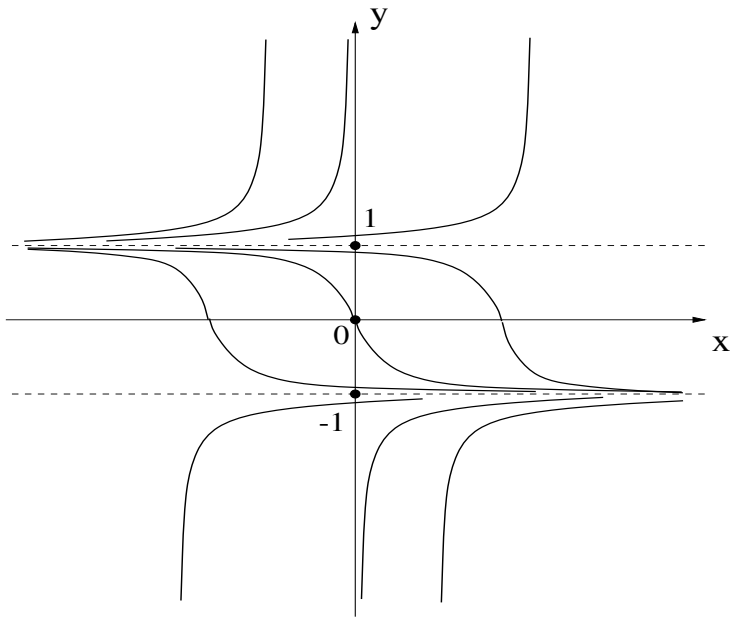


Fig. 3.8 Solutions for Example 3.9.

EXAMPLE 3.10. Find all the solutions of the DE

$$\frac{dy}{dx} = e^x \cos^2 y. \tag{3.94}$$

Solution: First, it is easy to see that

$$y = (2\,n + 1)\,\frac{\pi}{2} \quad (n = 0, \pm 1, \pm 2, \cdots) \tag{3.95}$$

are solutions of (3.94). When $y \neq (2\,n + 1)\,\frac{\pi}{2}$ $(n = 0, \pm 1, \pm 2, \cdots)$, for example, when $-\frac{\pi}{2} < y < \frac{\pi}{2}$, the solution of (3.94) is determined by

$$\int \frac{dy}{\cos^2 y} = \int e^x\,dx + C \qquad \text{i.e.,} \quad \tan y = e^x + C.$$

Hence,

$$y = \tan^{-1}(e^x + C), \qquad -\infty < x < +\infty. \tag{3.96}$$

Since the inverse trigonometric functions are multivalue functions, so for a given constant C, (3.96) actually represents a set of infinite solutions.

3.6.2 Exact first-order equations. If $f(x,\, y) = C$ (C is an arbitrary constant) is the general solution of an ODE, then

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy = 0 \tag{3.97}$$

and

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) \tag{3.98}$$

if function f is *smooth*; or more generally if f is *analytic*, then the conditions (3.97) and (3.98) are satisfied. So if we have an ODE given by

$$g(x,\, y)\,dx + h(x,\, y)\,dy = 0 \tag{3.99}$$

and

$$\frac{\partial g(x,\, y)}{\partial y} = \frac{\partial h(x,\, y)}{\partial x}, \tag{3.100}$$

then we expect there is a function $f(x,\, y)$ that produces the differential; and $\frac{\partial f}{\partial x} = g(x,\, y)$ and $\frac{\partial f}{\partial y} = h(x,\, y)$. The technique is to antidifferentiate g and h to find f. Actually,

$$f(x,\, y) = \int \frac{\partial f}{\partial x}\,dx + K(y) \qquad \text{or} \qquad f(x,\, y) = \int \frac{\partial f}{\partial y}\,dy + L(x). \tag{3.101}$$

We use the following example to illustrate the procedure of finding the solution of the DE (3.99).

EXAMPLE 3.11. Find all the solutions of the DE

$$(3\,x^2\,y + y^2 + 1)\,dx + (x^3 + 2\,x\,y - 1)\,dy = 0. \tag{3.102}$$

Solution: Here,

$$g(x,\,y) = 3\,x^2\,y + y^2 + 1 \qquad \text{and} \qquad h(x,\,y) = x^3 + 2\,x\,y - 1. \tag{3.103}$$

Obviously,

$$\frac{\partial g}{\partial y} = 3\,x^2 + 2\,y = \frac{\partial h}{\partial x}.$$

Therefore,

$$\begin{aligned} f(x,\,y) &= \int \frac{\partial f}{\partial x}\,dx + K(y) = \int g(x,\,y)\,dx + K(y) \\ &= \int (3\,x^2\,y + y^2 + 1)\,dx + K(y) \\ &= x^3\,y + x\,y^2 + x + K(y) \end{aligned} \tag{3.104}$$

and thus

$$\frac{\partial f}{\partial y} = x^3 + 2\,x\,y + \frac{dK}{dy} \tag{3.105}$$

which should be equal to $h(x,\,y)$. So, we have

$$\frac{dK}{dy} = -1 \qquad \Longrightarrow \qquad K(y) = -\,y. \tag{3.106}$$

Hence,

$$f(x,\,y) = x^3\,y + x\,y^2 + x - y. \tag{3.107}$$

Finally, $f(x,\,y) = C$ or

$$x^3\,y + x\,y^2 + x - y = C \tag{3.108}$$

represents all solutions of (3.102).

Sometimes, however, a given DE

$$g(x,\,y)\,dx + h(x,\,y)\,dy = 0$$

may do not satisfy $\dfrac{\partial g}{\partial y} = \dfrac{\partial h}{\partial x}$. In other words, it is not an exact first-order DE. But if we multiplying an appropriate function, $\mu(x,\, y) \neq 0$, to the above DE so that the modified DE

$$\mu(x,\, y)\, g(x,\, y)\, dx + \mu(x,\, y)\, h(x,\, y)\, dy = 0 \tag{3.109}$$

becomes exact, we call such a function $\mu(x,\, y)$ as *integrating factor*. How to find such a function $\mu(x,\, y)$? Since (3.109) is exact, so

$$\frac{\partial(\mu\, g)}{\partial y} = \frac{\partial(\mu\, h)}{\partial x}$$

or

$$\mu\, \frac{\partial g}{\partial y} + g\, \frac{\partial \mu}{\partial y} = \mu\, \frac{\partial h}{\partial x} + h\, \frac{\partial \mu}{\partial x}$$

which can be written as

$$\frac{1}{\mu}\left(h\, \frac{\partial \mu}{\partial x} - g\, \frac{\partial \mu}{\partial y}\right) = \frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}$$

that is,

$$h\, \frac{\partial \ln \mu}{\partial x} - g\, \frac{\partial \ln \mu}{\partial y} = \frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}. \tag{3.110}$$

This is a partial differential equation (PDE). In general, solving for a PDE is more difficulty than solving for an ODE. However, in some special cases, we may easily to find a solution of (3.110). Having find the integrating factor $\mu(x,\, y)$, we can use the procedure of exact first-order ODE to find all the solutions of (3.109).

We consider two special cases:

(I) μ is independent of y, then $\dfrac{\partial \ln \mu}{\partial y} = 0$ and (3.110) becomes

$$\frac{d \ln \mu}{dx} = \frac{1}{h}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right). \tag{3.111}$$

Since the left-hand side depends upon x only, so the necessary and sufficient conditions for (3.111) to have a solution for μ is that

$$\frac{1}{h}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right)$$

is independent of y. Then,

$$\mu(x) = \exp\left[\int \frac{1}{h}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right) dx\right]. \tag{3.112}$$

(II) Similarly, if

$$\frac{1}{g}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right)$$

is independent of x, then,

$$\mu(y) = \exp\left[-\int \frac{1}{g}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right) dy\right]. \tag{3.113}$$

EXAMPLE 3.12. Find all the solutions of the DE

$$\left(2\,x\,y + x^2\,y + \frac{y^3}{3}\right) dx + (x^2 + y^2)\, dy = 0. \tag{3.114}$$

Solution: For this example,

$$g(x,\, y) = 2\,x\,y + x^2\,y + \frac{y^2}{3} \qquad \text{and} \qquad h(x,\, y) = x^2 + y^2 \tag{3.115}$$

and so

$$\frac{\partial g}{\partial y} = 2\,x + x^2 + y^2 \neq \frac{\partial h}{\partial x} = 2\,x.$$

Thus (3.114) is not exact. However,

$$\frac{1}{h}\left(\frac{\partial g}{\partial y} - \frac{\partial h}{\partial x}\right) = \frac{1}{x^2 + y^2}\,(2\,x + x^2 + y^2 - 2\,x) = 1$$

which does not depend upon x, so

$$\mu(x) = e^{\int 1\, dx} = e^x. \tag{3.116}$$

Then the DE

$$e^x \left(2\,x\,y + x^2\,y + \frac{y^3}{3}\right) dx + e^x\,(x^2 + y^2)\, dy = 0 \tag{3.117}$$

is now exact. We want to find $f(x,\, y)$ such that

$$\frac{\partial f}{\partial x} = e^x \left(2\,x\,y + x^2\,y + \frac{y^3}{3}\right) \qquad \text{and} \qquad \frac{\partial f}{\partial y} = e^x\,(x^2 + y^2).$$

Thus,

$$\begin{aligned} f(x,\,y) &= \int \frac{\partial f}{\partial x}\,dx + K(y) \\ &= \int e^x \left(2\,x\,y + x^2\,y + \frac{y^3}{3}\right) + K(y) \\ &= y\,x^2\,e^x + \frac{1}{3}\,y^3\,e^x + K(y) \end{aligned} \tag{3.118}$$

and so

$$\frac{\partial f}{\partial y} = x^2\,e^x + y^2\,e^x + \frac{dK}{dy} = e^x\,(x^2 + y^2). \tag{3.119}$$

Hence, we have

$$\frac{dK}{dy} = 0 \qquad \Longrightarrow \qquad K(y) \equiv K(y_0) \qquad (K(y_0) \text{ is a constant}). \tag{3.120}$$

Choosing $K(y_0) = 0$ yields

$$f(x,\,y) = y\,e^x \left(x^2 + \frac{1}{3}\,y^2\right). \tag{3.121}$$

Finally,

$$y\,e^x \left(x^2 + \frac{1}{3}\,y^2\right) = C \tag{3.122}$$

represents all the solutions of (3.114).

3.6.3 Homogeneous type ODE. Consider the DE

$$\frac{dy}{dx} = f(x,\,y). \tag{3.123}$$

If $f(x,\,y)$ satisfies

$$f(\tau\,x,\,\tau\,y) = f(x,\,y) \qquad \forall \;\; \tau \neq 0 \tag{3.124}$$

then we call the DE (3.123) a *homogeneous type* ODE. This kind of DEs can be transformed to separable equations. In fact, by introducing a new variable v such that $y = v\,x$, we have

$$\frac{dy}{dx} = v + x\,\frac{dv}{dx} = f(x,\,v\,x) = f(1,\,v) \qquad \Longrightarrow \qquad x\,\frac{dv}{dx} = f(1,\,v) - v \tag{3.125}$$

which is a separable DE.

EXAMPLE 3.13. Find all the solutions of the DE

$$\frac{dy}{dx} = \frac{3\,y^2 - x^2}{2\,x\,y}. \tag{3.126}$$

Solution: This is a homogeneous type ODE, so let $y = v\,x$, then (3.126) becomes

$$v + x\,\frac{dv}{dx} = \frac{3\,v^2\,x^2 - x^2}{2\,v\,x^2} = \frac{3\,v^2 - 1}{2\,v} \tag{3.127}$$

or

$$x\,\frac{dv}{dx} = \frac{v^2 - 1}{2\,v}$$

which is separable. We rewrite it as

$$\frac{2\,v\,dv}{v^2 - 1} = \frac{dx}{x} \qquad \text{i.e.,} \quad \frac{d(v^2 - 1)}{v^2 - 1} = \frac{dx}{x} \tag{3.128}$$

which gives the solution

$$\ln\left|v^2 - 1\right| = \ln|x| + C_1$$

$$\Longrightarrow$$

$$\left|v^2 - 1\right| = e^{C_1 + \ln|x|} = e^{C_1}\ |x| = C_2\ |x| \qquad (C_2 = e^{C_1} > 0)$$

$$\Longrightarrow$$

$$v^2 - 1 = \pm\,C_2\,x = C\,x \qquad (C = \pm\,C_2 \neq 0) \tag{3.129}$$

Substituting v into $y = v\,x$ results in

$$\frac{y^2}{x^2} - 1 = C\,x \qquad \text{or} \qquad y^2 = C\,x^3 + x^2 \tag{3.130}$$

which is the general solution of (3.126).

Some DEs can be transformed to homogeneous type DEs, for example, the DE

$$\frac{dy}{dx} = \frac{a_1\,x + b_1\,y + c_1}{a_2\,x + b_2\,y + c_2} \tag{3.131}$$

is one of these type DEs, where $a_i,\ b_i,\ c_i,\ i = 1,\ 2$ are given constants.

If $c_1 = c_2 = 0$, then $f(x,\ y) = \dfrac{a_1\,x + b_1\,y}{a_2\,x + b_2\,y}$, so (3.131) is a homogeneous type DE. If c_1 and c_2 are not both equal to zero, then (3.131) is not a homogeneous type DE. However, we may introduce a transformation

$$x = \xi + h, \quad y = \eta + k \tag{3.132}$$

where ξ and η are new variables, and h and k are coefficients to be determined, into transform (3.131) to obtain

$$\frac{d\eta}{d\xi} = \frac{a_1\,\xi + b_1\,\eta + a_1\,h + b_1\,k + c_1}{a_2\,\xi + b_2\,\eta + a_2\,h + b_2\,k + c_2}. \tag{3.133}$$

If we can choose h and k such that

$$a_1\,h + b_1\,k + c_1 = 0 \qquad \text{and} \qquad a_2\,h + b_2\,k + c_2 = 0; \tag{3.134}$$

then (3.131) becomes a homogeneous type DE

$$\frac{d\eta}{d\xi} = \frac{a_1\,\xi + b_1\,\eta}{a_2\,\xi + b_2\,\eta}. \tag{3.135}$$

Now how to choose h and k to satisfy (3.134)? Since (3.134) represent two linear algebraic equations about h and k and we know from linear algebra that the two equations determine a unique solution for h and k if the determinant $a_2\,b_1 - a_1\,b_2 \neq 0$.

If $a_2\,b_1 - a_1\,b_2 = 0$, (3.134) in general has no solution for h and k and thus the above procedure fails. However, in this case, we can still transform the original DE (3.131) to a separable equation, which is shown as follows. If $b_2 \neq 0$, then $\dfrac{a_1}{a_2} = \dfrac{b_1}{b_2} = \lambda$, so (3.131) becomes

$$\frac{dy}{dx} = \frac{\lambda\,(a_2\,x + b_2\,y) + c_1}{a_2\,x + b_2\,y + c_2} \tag{3.136}$$

which can be transformed into the following separable equation

$$\frac{dz}{dx} = a_2 + b_2\,\frac{\lambda\,z + c_1}{z + c_2} \tag{3.137}$$

by introducing $z = a_2\,x + b_2\,y$. If $b_2 = 0$, then $a_2\,b_1 - a_1\,b_2 = a_2\,b_1$ and there are only two possibilities: either $a_2 = 0$ or $b_1 = 0$. If $a_2 = 0$, (3.131) is now in the form

$$\frac{dy}{dx} = a_1\,x + b_1\,y + c_1. \tag{3.138}$$

By letting $z = a_1\,x + b_1\,y$, we can transform (3.138) to the separable equation

$$\frac{dz}{dx} = a_1 + b_1\,(z + c). \tag{3.139}$$

While if $b_1 = 0$, then (3.131) becomes

$$\frac{dy}{dx} = \frac{a_1\, x + c_1}{a_2\, x + c_2} \tag{3.140}$$

which is already in separable form.

From the above analysis, we conclude that (3.131) can always be transformed to a separable equation and therefore we can find the solution of the DE. The equation (3.134) can be considered as a translation, which moves the origin to $(h,\, k)$ which is actually the intersection point of the two lines

$$a_1\, x + b_1\, y + c_1 = 0 \qquad \text{and} \qquad a_2\, x + b_2\, y + c_2 = 0.$$

The procedure described above can be applied to the more general DE

$$\frac{dy}{dx} = f\left(\frac{a_1\, x + b_1\, y + c_1}{a_2\, x + b_2\, y + c_2}\right). \tag{3.141}$$

EXAMPLE 3.14. Solve for the DE

$$\frac{dy}{dx} = 2\left(\frac{y+2}{x+y-1}\right)^2. \tag{3.142}$$

Solution: In this example, $a_1 = 0,\ b_1 = 1,\ c_1 = 2$ and $a_2 = 1,\ b_2 = 1,\ c_2 = -1$. Thus, $a_2\, b_1 - a_1\, b_2 = 1 - 0 = 1 \neq 0$. So we can choose h and h such that

$$k + 2 = 0 \qquad \text{and} \qquad h + k - 1 = 0 \tag{3.143}$$

which yields the solution $h = 3,\ k = -2$. Next, by using the transformation

$$x = \xi + 3, \quad y = \eta - 2 \tag{3.144}$$

one obtains

$$\frac{d\eta}{d\xi} = 2\left(\frac{\eta}{\xi + \eta}\right)^2. \tag{3.145}$$

Then introducing another transformation $\eta = v\,\xi$ to transform (3.145) to

$$v + \xi\,\frac{dv}{d\xi} = 2\left(\frac{v}{1+v}\right)^2$$

or

$$\xi \frac{dv}{d\xi} = -\frac{v\,(1+v^2)}{(1+v)^2}. \tag{3.146}$$

The solution of (3.146) is either $v = 0$, i.e., $\eta = 0$ or is given by the integration

$$\int \left(\frac{1}{v} + \frac{2}{1+v^2}\right) dv = -\int \frac{d\xi}{\xi} + C' \tag{3.147}$$

which is

$$\ln|v| + 2\,\tan^{-1} v = -\ln|\xi| + C'$$

$$\Longrightarrow$$

$$\ln|\xi\, v| = -2\,\tan^{-1} v + C'$$

$$\Longrightarrow$$

$$\eta = C\, e^{-2\,\tan^{-1}\frac{\eta}{\xi}} \qquad (C = \pm e^{C'} \neq 0). \tag{3.148}$$

Combining $\eta = 0$ and (3.148) yields the solution

$$\eta = C\, e^{-2\,\tan^{-1}\frac{\eta}{\xi}}$$

where C is an arbitrary constant. Hence the solution to the original DE (3.142) is

$$y + 2 = C\, e^{-2\,\tan^{-1}\frac{y+2}{x-3}}. \tag{3.149}$$

3.6.4 Transformations. Some nonlinear DEs can be transformed to linear DEs by an appropriate transformation.

1) **Bernoulli equation**

$$\frac{dy}{dx} + P(x)\, y = Q(x)\, y^{\alpha}. \tag{3.150}$$

a) $\alpha = 0$, (3.150) is a linear DE;

b) $\alpha = 1$, (3.150) is a separable equation;

When α is a natural number, $y = 0$ is a solution of (3.150). Suppose $\alpha \neq 0,\ 1$ and $y \neq 0$, introduce

$$v = y^{-\alpha+1} \tag{3.151}$$

into (3.150) to obtain

$$\frac{y^\alpha}{1-\alpha}\frac{dv}{dx} + P(x)\,y^\alpha\,v = Q(x)\,y^\alpha$$

$$\Longrightarrow$$

$$\frac{dv}{dx} + (1-\alpha)\,P(x)\,v = (1-\alpha)\,Q(x) \tag{3.152}$$

which is a *linear* ODE of v. We can first solve out v and substitute it into (3.151) to obtain solution y for the original DE (3.150).

1) **Riccati equation**

$$\frac{dy}{dx} = P(x)\,y^2 + Q(x)\,y + R(x). \tag{3.153}$$

The right-hand side of (3.153) is a quadratic function of y. Except for some special cases, we cannot find analytic solutions for (3.153). However, if we know a solution of DE (3.153) (e.g., by inspection), then we can find all solutions of the DE. This can be achieved as follows. Suppose one solution is given by

$$y = \phi(x). \tag{3.154}$$

Then by using the transformation

$$y = v + \phi(x) \tag{3.155}$$

we obtain

$$\begin{aligned} & \frac{dv}{dx} + \frac{d\phi(x)}{dx} \\ = \;& P(x)\,[\,v+\phi(x)\,]^2 + Q(x)\,[\,v+\phi(x)\,] + R(x) \\ = \;& P(x)\,v^2 + [\,2\,P(x)\,\phi(x) + Q(x)\,]\,v + P(x)\,\phi(x)^2 + Q(x)\,\phi(x) + R(x). \end{aligned} \tag{3.156}$$

Since $\phi(x)$ is a solution of (3.153), so

$$\frac{d\phi(x)}{dx} = P(x)\,\phi(x)^2 + Q(x)\,\phi(x) + R(x) \tag{3.157}$$

and thus (3.156) becomes

$$\frac{dv}{dx} = P(x)\,v^2 + [\,2\,P(x)\,\phi(x) + Q(x)\,]\,v \tag{3.158}$$

which is a *Bernoulli equation.*

EXAMPLE 3.15. Find all solutions of the DE

$$\frac{dy}{dx} = y^2 - \frac{2}{x^2}. \tag{3.159}$$

Solution: This is a Riccati equation with $P(x) = 1$, $Q(x) = 0$ and $R(x) = -\frac{2}{x^2}$. It is easy to check that $y = \frac{1}{x}$ is a solution:

$$\text{LHS} = -\frac{1}{x^2} = \text{RHS} = \frac{1}{x^2} - \frac{2}{x^2} = -\frac{1}{x^2}.$$

So letting $y = v + \frac{1}{x}$ and substituting it into (3.159) yields

$$\frac{dv}{dx} = v^2 + \frac{2}{x}\, v \tag{3.160}$$

which is a Bernoulli equation. It has a solution

$$v = 0. \tag{3.161}$$

When $v \neq 0$, dividing both sides of (3.160) by v^2 results in

$$v^{-2}\, \frac{dv}{dx} = 1 + \frac{2}{x}\, v^{-1}. \tag{3.162}$$

Next, let

$$z = \frac{1}{v} \tag{3.163}$$

and then (3.162) becomes

$$\frac{dz}{dx} = -\, v^{-2}\, \frac{dv}{dx} = -1 - \frac{2}{x}\, z. \tag{3.164}$$

This is a linear DE and its solution is given by (see equation (2.95))

$$\begin{aligned} z(x) &= e^{-\int \frac{2}{x}\, dx} \left(C' - \int e^{\int \frac{2}{x}\, dx}\, dx \right) \\ &= x^{-2} \left(C' - \frac{1}{3}\, x^3 \right). \end{aligned} \tag{3.165}$$

So, the solutions of the original ODE (3.159) are

$$y = \frac{1}{x} \qquad (\text{from } \ v = 0) \tag{3.166}$$

and (from $v \neq 0$)

$$y = v + \frac{1}{x} = \frac{1}{z} + \frac{1}{x} = \frac{x^2}{C' - x^3/3} + \frac{1}{x}$$
$$= \frac{3\,x^2}{C - x^3} + \frac{1}{x} \qquad (C = 3\,C'). \tag{3.167}$$

Equation (3.167) represents a family of solutions of (3.159). Since (3.66) does not includes arbitrary constants, let's check if it is a singular solution of (3.159). Eliminating C from the following two equations:

$$\phi(x,\, y,\, C) = y - \frac{3\,x^2}{C - x^3} - \frac{1}{x} = 0$$
$$\frac{\partial \phi}{\partial C} = \frac{-\,3\,x^2}{(C - x^3)^2} = 0 \tag{3.168}$$

yields $y = \dfrac{1}{x}$. Hence, $y = \dfrac{1}{x}$ is a singular solution of (3.159).

3) **Clairaut equation**

$$y = x\,\frac{dy}{dx} + f\left(\frac{dy}{dx}\right). \tag{3.169}$$

Let $\dfrac{dy}{dx} = p$, then

$$y = x\,p + f(p). \tag{3.170}$$

Suppose $f(p)$ is not identically constant. Differentiating (3.170) with respect to x results in

$$p = p + [\,x + f'(p)\,]\,\frac{dp}{dx}, \tag{3.171}$$

that is,

$$\frac{dp}{dx} = 0 \qquad \text{or} \qquad x + f'(p) = 0. \tag{3.172}$$

From the first equation $\dfrac{dp}{dx} = 0$, we find $p = C$, and then substitute it into (3.170) to obtain

$$y = C\,x + f(C) \tag{3.173}$$

which is a family of straight lines. (3.173) can be directly obtained from the original DE (3.169) by substituting C for $\dfrac{dy}{dx}$. The envelope of (3.173) is determined by the C-discriminant equation solved from

$$\phi(x,\, y,\, C) = y - C\,x - f(C) = 0$$
$$\frac{\partial \phi}{\partial C} = x + f'(C) = 0$$

in a parametric form

$$\begin{aligned} x &= -f'(C) \\ y &= -C f'(C) + f(C). \end{aligned} \tag{3.174}$$

On the other hand, from the second equation of (3.172) and (3.173), we find

$$\begin{aligned} x &= -f'(p) \\ y &= -p f'(p) + f(p) \end{aligned} \tag{3.175}$$

which is p-discriminant equation. It has the same form as the C-discriminant equation (3.174). Therefore, (3.175) represents a singular solution of (3.169) in a parametric form. This can be also directly verified as follows. Thinking p as a parameter, then the slop of the tangent line of the curve denoted by the equation (3.175) is

$$\frac{dy}{dx} = \frac{dy/dp}{dx/dp} = \frac{-p f''(p)}{-f''(p)} = p \qquad (\text{suppose } f''(p) \neq 0). \tag{3.176}$$

Then substituting $x = -f'(p)$ and $\dfrac{dy}{dx} = p$ into the second equation of (3.175) yields

$$y = x \frac{dy}{dx} + f\left(\frac{dy}{dx}\right)$$

which is actually the original DE (3.169).

Therefore, the Clairaut equation (3.169) has the family of solutions (3.173) and the singular solution (3.175).

EXAMPLE 3.16. Solve the DE

$$y = \frac{dy}{dx} x + \frac{m}{\dfrac{dy}{dx}} \tag{3.177}$$

where m is a constant.

Solution: This is a Clairaut equation. Let $\dfrac{dy}{dx} = p$, then (3.177) becomes

$$y = p x + \frac{m}{p}. \tag{3.178}$$

Differentiating (3.178) with respect to x gives us

$$\begin{aligned} & \frac{dy}{dx} = p + x\,\frac{dp}{dx} - \frac{m}{p^2}\,\frac{dp}{dx} \\ \Longrightarrow & \\ & \left(x - \frac{m}{p^2}\right)\frac{dp}{dx} = 0 \\ \Longrightarrow & \\ & \frac{dp}{dx} = 0 \qquad \text{or} \qquad x = \frac{m}{p^2}. \end{aligned} \tag{3.179}$$

Thus, we have a family of solutions obtained by substituting $p = C$ (due to $\dfrac{dp}{dx} = 0$) into (3.178):

$$y = C\,x + \frac{m}{C}. \tag{3.180}$$

Another solution is obtained from $x = \dfrac{m}{p^2}$ or $p^2 = \dfrac{m}{x}$. Substituting $p^2 = \dfrac{m}{x}$ into (3.178) leads to

$$\begin{aligned} y^2 &= \left(p\,x + \frac{m}{p}\right)^2 \\ &= p^2\,x^2 + \frac{m^2}{p^2} + 2\,x\,m \\ &= \frac{m}{x}\,x^2 + \frac{m^2}{m/x} + 2\,x\,m \\ &= 4\,m\,x. \end{aligned} \tag{3.181}$$

We can verify that (3.181) is a singular solution of (3.177). Since eliminating C from

$$\begin{aligned} \phi(x,\,y,\,C) &= y - C\,x - \frac{m}{C} = 0 \\ \frac{\partial \phi}{\partial C} &= -\,x + \frac{m}{C^2} = 0 \end{aligned} \tag{3.182}$$

results in $y^2 = 4\,m\,x$. So $y^2 = 4\,m\,x$ is the singular solution of (3.177) and it is the envelope of the family of solutions (3.180).

3.7 Some High-Order NLODEs

In this section, we consider two kinds of high-order nonlinear ordinary differential equations which can be solved by using some appropriate substitutions to reduce the order of the DEs.

3.7.1 Dependent variable missing. Suppose the DE is given by

$$F\left(x,\, \frac{d^k y}{dx},\, \cdots,\, \frac{d^n y}{dx^n}\right) = 0. \tag{3.183}$$

Let $v = \dfrac{d^k y}{dx^k}$, we obtain, from (3.183), a $(n-k)$-order DE

$$F\left(x,\, v,\, \frac{dv}{dx},\, \cdots,\, \frac{d^{n-k} v}{dx^{n-k}}\right) = 0. \tag{3.184}$$

If we find a solution of (3.184): $v = \phi(x)$, then the solution of the DE

$$\frac{d^k y}{dx^k} = \phi(x) \tag{3.185}$$

is a solution of the original DE (3.183). So in order to find the solution of (3.183), we need only to solve the lower order DEs (3.184) and (3.185).

EXAMPLE 3.17. Solve the DE

$$\frac{d^5 y}{dx^5} - \frac{1}{x}\,\frac{d^4 y}{dx^4} = 0. \tag{3.186}$$

Solution: Introduce $v = \dfrac{d^4 y}{dx^4}$ into (3.186) to obtain

$$\frac{dv}{dx} - \frac{1}{x}\, v = 0 \tag{3.187}$$

which is a first-order separable equation and its solution is

$$v = C\, x = 120\, C_1\, x \qquad (C = 120\, C_1). \tag{3.188}$$

Thus integrating the DE

$$\frac{d^4 y}{dx^4} = 120\, C_1\, x \tag{3.189}$$

yields

$$y = C_1\, x^5 + C_2\, x^3 + C_3\, x^2 + C_4\, x + C_5 \tag{3.190}$$

where $C_i,\ i = 1,\, 2,\, \cdots,\, 5$ are arbitrary constants.

EXAMPLE 3.18. Find all the solutions of the DE

$$x\, y'' - (y')^3 - y' = 0. \tag{3.191}$$

Solution: Let $p = \dfrac{dy}{dx}$, then (3.191) becomes

$$x\,\frac{dp}{dx} - p^3 - p = 0 \tag{3.192}$$

which is a first-order equation about p and is separable. First note that $p = 0$ or $y = C$ is a solution of (3.191). Other solutions can be found as follows.

$$\frac{dp}{p\,(p^2+1)} = \frac{dx}{x}$$

$$\Longrightarrow$$

$$\int \left(\frac{1}{p} - \frac{p}{p^2+1}\right) dp = \int \frac{dx}{x} \qquad \text{(partial fractions)}$$

$$\Longrightarrow$$

$$\ln|p| - \frac{1}{2}\ln(p^2+1) = \ln|x| + C'$$

$$\Longrightarrow$$

$$\ln \frac{|p|}{\sqrt{p^2+1}} = \ln|x| + C'$$

$$\Longrightarrow$$

$$\frac{|p|}{\sqrt{p^2+1}} = e^{C'}\,|x| = C''\,|x| \qquad (C'' = e^{C'} > 0)$$

$$\Longrightarrow$$

$$\frac{p^2}{p^2+1} = (C'')^2\,x^2$$

$$\Longrightarrow$$

$$p^2 = \frac{(C'')^2\,x^2}{1-(C'')^2\,x^2} = \frac{x^2}{1/(C'')^2 - x^2} = \frac{x^2}{C_1^2 - x^2} \qquad (C_1 = \frac{1}{C''} > 0)$$

$$\Longrightarrow$$

$$p = \frac{\pm\,x}{\sqrt{C_1^2 - x^2}} \tag{3.193}$$

Thus $\dfrac{dy}{dx} = p = \dfrac{\pm\,x}{\sqrt{C_1^2 - x^2}}$, and integrating this equation yields

$$\begin{aligned} y &= \int \frac{\pm\,x\,dx}{\sqrt{C_1^2 - x^2}} \qquad (\text{let } \; u = C_1^2 - x^2 \;\text{ and so }\; du = -\,2\,x\,dx) \\ &= \pm\,\frac{1}{2}\int \frac{du}{\sqrt{u}} \end{aligned}$$

$$\begin{aligned} &= \pm\sqrt{u} + C_2 \\ &= C_2 \pm \sqrt{C_1^2 - x^2}, \end{aligned} \tag{3.194}$$

and finally we have the solution

$$x^2 + (y - C_2)^2 = C_1^2 \qquad (C_1 > 0) \tag{3.195}$$

for the original DE (3.191). Remember that $y = C$ is also a solution of (3.191).

3.7.2 Independent variable missing. In this case the DE can be written as

$$F\left(y,\, y',\, \cdots,\, y^{(n)}\right) = 0. \tag{3.196}$$

Many problems of mechanics can be described by (3.196). For example, the pendulum model (see Example 2.3). In this case, again let $p = y' = \dfrac{dy}{dx}$, but now consider y as independent variable and want to find the relation between p and y. Obviously,

$$\begin{aligned} y' &= p \\ y'' &= \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{dp}{dx} = \frac{dp}{dy}\frac{dy}{dx} = p\frac{dp}{dy} \\ y''' &= \frac{d}{dx}\left(\frac{d^2y}{dx^2}\right) = \frac{d}{dx}\left(p\frac{dp}{dy}\right) \\ &= p\frac{d}{dy}\left(p\frac{dp}{dy}\right) = p\left(\frac{dp}{dy}\right)^2 + p^2\frac{d^2p}{dy^2} \\ &\cdots \end{aligned} \tag{3.197}$$

It can be proved by mathematics induction that $\dfrac{d^k y}{dx^k}$ can be expressed in terms of p, $\dfrac{dp}{dy}$, $\cdots$, $\dfrac{d^{k-1}y}{dy^{k-1}}$. Substituting these expressions into the DE (3.196) yields $(n-1)$-order DE

$$F\left(y,\, p,\, p\frac{dp}{dy},\, p\left(\frac{dp}{dy}\right)^2 + p^2\frac{d^2p}{dy^2},\, \cdots\right) = 0. \tag{3.198}$$

EXAMPLE 3.19. Solve the DE

$$y\,y'' + (y')^2 + 1 = 0. \tag{3.199}$$

Solution: Let $y' = \dfrac{dy}{dx} = p$, so $y'' = p\dfrac{dp}{dy}$, and then (3.199) becomes

$$y\,p\frac{dp}{dy} + p^2 + 1 = 0 \tag{3.200}$$

which is a separable first-order ODE. We can find its solution as follows. First note that $y = C$ is not solution of (3.199). So suppose that y is not a constant; i.e., $p = \dfrac{dy}{dx} \neq 0$, then

$$
\begin{aligned}
& y\,p\,\frac{dp}{dy} + p^2 + 1 = 0 \\
\Longrightarrow \quad & \\
& \frac{p\,dp}{p^2+1} + \frac{dy}{y} = 0 \\
\Longrightarrow \quad & \\
& \frac{1}{2}\,d\,\ln(p^2+1) + d\ln|y| = 0 \\
\Longrightarrow \quad & \\
& d\,\ln(p^2+1) + d\ln y^2 = 0 \\
\Longrightarrow \quad & \\
& d\,\ln\left[\,(p^2+1)\,y^2\,\right] = 0 \\
\Longrightarrow \quad & \\
& (p^2+1)\,y^2 = e^C = C_1^2 \qquad (C_1^2 = e^C > 0) \\
\Longrightarrow \quad & \\
& p^2 = \frac{C_1^2}{y^2} - 1 \\
\Longrightarrow \quad & \\
& p = \pm\sqrt{\frac{C_1^2}{y^2} - 1} = \pm\frac{1}{y}\sqrt{C_1^2 - y^2} \qquad (3.201)
\end{aligned}
$$

and thus

$$
\begin{aligned}
& \frac{dy}{dx} = \pm\frac{1}{y}\sqrt{C_1^2 - y^2} \\
\Longrightarrow \quad & \\
& \frac{y\,dy}{\sqrt{C_1^2 - y^2}} = \pm\,dx \\
\Longrightarrow \quad & \\
& -\frac{1}{2}\,\frac{d(C_1^2 - y^2)}{\sqrt{C_1^2 - y^2}} = \pm\,dx \\
\Longrightarrow \quad &
\end{aligned}
$$

$$\sqrt{C_1^2 - y^2} = \pm (x - C_2)$$

$$\Longrightarrow$$

$$C_1^2 - y^2 = (x - C_2)^2$$

$$\Longrightarrow$$

$$(x - C_2)^2 + y^2 = C_1^2. \tag{3.202}$$

Chapter 4

BIFURCATIONS IN DISCRETE MAP

A dynamical system with physical interest typically has several independent parameters which appear in the defining systems of equations. Thus, it is important to understand the qualitative properties of the system as the parameters vary. A change in the qualitative properties could mean a change in stability of the original system and thus the system must assume a state different from the original one. These changes are often called *bifurcations* and the values of the parameters at which a sudden "jump" takes place are called *bifurcation values*. For example. consider the buckling of a beam. If a small weight is placed on top of the beam in Fig. 4.1, the beam can support the load and remain vertical. But if the load is too heavy, the vertical position becomes unstable, and the beam may buckle. Here the weight plays the role of (control) parameter, and the deflection of the beam from vertical plays the role of the dynamical variable.

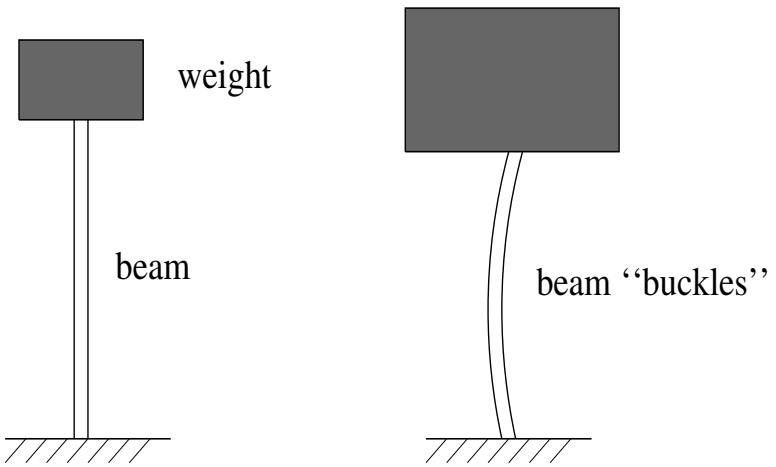


Fig. 4.1 Buckling of a beam.

Other examples can be found from different disciplines:

Physics: water $\longrightarrow$ steam (boiling), ice $\longrightarrow$ water (melting);

Solid mechanics: buckling, flutter;

Fluid mechanics: Laminar flow $\longrightarrow$ turbulence;

Chemistry: match $\longrightarrow$ fire;

Aeronautics: aircraft at high-angle-of-attack $\longrightarrow$ oscillation;

Electrical circuits: self-excited oscillation;

Astronomy: giant stars collapse;

Laser: ordinary light $\longrightarrow$ coherent light

Ecology: density of population undergoes a saccadic leap;

Financial: boar stock market $\longrightarrow$ bear stock market;

Economics: economic crisis;

Social science: peace $\longrightarrow$ war;

Daily life: birth, death, earthquake;

etc.

One of main goals of this course it to help you bifurcation phenomena. We study the bifurcations for discrete maps in this chapter and for continuous differential equations in next chapter. We start by considering a simple example – logistic system.

4.1 Logistic Equation

Consider the following first-order ODE

$$\frac{dP(t)}{dt} = k\,P(t) \tag{4.1}$$

where $P(t)$ denotes population of some organism at time t. This would be realistic if growth were determined exclusively by an existing population. It has, of course, a solution

$$P(t) = P_0\,e^{kt} \tag{4.2}$$

where k could represent the growth ($k > 0$) due to reproduction, or the result of death ($k < 0$). This picture allows for only unchecked growth or extinction, nothing else.

A more realistic model would establish a fixed population L after some initial transient behavior. The model is described by

$$\frac{dP(t)}{dt} = k\,P(t)\,(L - P(t))\,. \tag{4.3}$$

The factor installed, is not unique, but it causes the solutions to have the following interesting properties for positive k. Suppose $k > 0$, then

$$\begin{array}{lllll} \text{for} & P = L & \Longrightarrow & \dfrac{dP}{dt} = 0 & \text{population does not change;} \\ & P > L & \Longrightarrow & \dfrac{dP}{dt} < 0 & \text{population decreases;} \\ & P < L & \Longrightarrow & \dfrac{dP}{dt} > 0 & \text{population increases.} \end{array} \tag{4.4}$$

This equation is known as the *logistic equation*† We see that no change in the solution arises when $P = L$. We call $P = L$ a *fixed point* of the equation. Many authors use the terms: *equilibrium solution, stationary point, singular point, critical point, steady state*, for the same purpose. We'll use whichever is most suitable, given that the meaning is largely selfevident. (When we get to phase space, we'll see that point is really meaningful.) We also see that this fixed point is a *stable fixed point*, in that "pushing" (disturbance, perturbation) the population P off of L leads to a return to L. This is not the case for the fixed point $P = 0$. In that case, $\dfrac{dP}{dt} = 0$, and no change can take place too, however, if the population is "pushed" off of zero (i.e., something is added to nothing) it "moves" away from 0 because $\dfrac{dP}{dt}$ becomes positive. We say that $P = 0$ is an *unstable fixed point.* (We'll discuss stability more later.) This idea is that $P = L$ is the ultimate outcome of any starting value of P; i.e., $P \neq P_0$, $P_0 > 0$. See Fig. 4.1. We say $P = L$ "attracts" the solutions other than $P = L$ (for $P_0 > 0$). $P = L$ is known as an *attractor.* Loosely speaking, an attractor is a set to which all neighboring trajectories (solutions) converge.

We have learned a great deal about the solutions of

$$\frac{dP}{dt} = k\,P\,(L - P)$$

†One can alternatively deduce this equation for the spread of a dicease. It is used in epidemiological modelling, and actuarial science

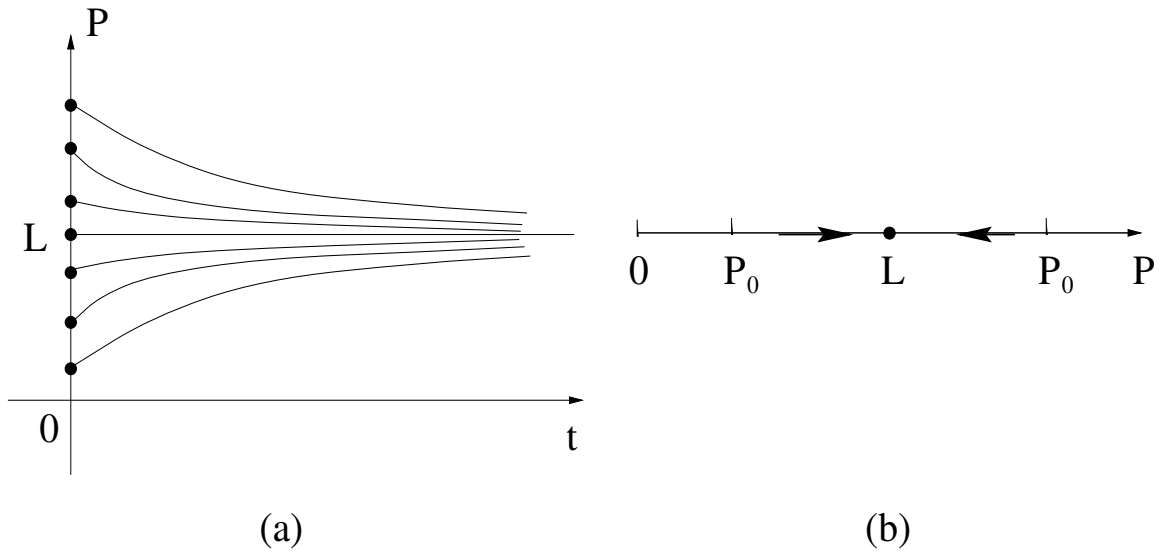


Fig. 4.2 Solutions of the logistic equation (4.3): (a) time history; (b) phase space.

without actually solving this equation with direct analytical techniques!

We have already seen that solutions, generally speaking, for nonlinear ODEs are *not* always available by direct analytical techniques, for example, Riccati Equations. Nevertheless, in this case, we can use the results given in the previous chapter to find an explicit form for the solutions of the logistic equation. Since at the fixed points $P = 0$ or $P = L$, $\dfrac{dP}{dt} = 0$. So in the following we assume $P > 0$ and $P \neq L$. Then the solution of the logistic equation can be obtained as follows. First note that

$$\frac{1}{P\,(L-P)} = \frac{A}{P} + \frac{B}{L-P} \qquad \text{(partial fraction)}$$

$$\Longrightarrow$$

$$1 = A\,(L-P) + B\,P = (B-A)\,P + A\,L$$

$$\Longrightarrow$$

$$B - A = 0 \qquad \text{and} \qquad A\,L = 1$$

$$\Longrightarrow$$

$$A = B = \frac{1}{L}. \tag{4.5}$$

Therefore,

$$\begin{aligned}
&\frac{dP}{dt} = k\,P\,(L-P) \\
\Longrightarrow \quad &\\
&\frac{dP}{P\,(L-P)} = k\,dt \\
\Longrightarrow \quad &\\
&\frac{1}{L}\left(\frac{1}{P}+\frac{1}{L-P}\right)dp = k\,dt \\
\Longrightarrow \quad &\\
&\int\left(\frac{1}{P}+\frac{1}{L-P}\right)dp \int L\,k\,dt \\
\Longrightarrow \quad &\\
&\ln\left|\frac{P}{L-P}\right| = L\,k\,t + C' \\
\Longrightarrow \quad &\\
&\frac{P}{L-P} = \pm\, e^{C'+L\,k\,t} = \pm\, e^{C'}\, e^{L\,k\,t} = C\, e^{L\,k\,t} \qquad (C = \pm\, e^{C'} \neq 0).
\end{aligned} \tag{4.6}$$

Suppose when $t = t_0,\ P = P_0$, then

$$\frac{P_0}{L-P_0} = C\, e^{L\,k\,t_0} \qquad \Longrightarrow \qquad C = \frac{P_0}{L-P_0}\, e^{-\,L\,k\,t_0};$$

and thus (4.6) becomes

$$\frac{P}{L-P} = \frac{P_0}{L-P_0}\, e^{L\,k\,(t-t_0)}. \tag{4.7}$$

Solving for P from (4.7) yields

$$\begin{aligned}
P &= \frac{L\,P_0\, e^{L\,k\,(t-t_0)}}{L - P_0 + P_0\, e^{L\,k\,(t-t_0)}} \\
&= \frac{L\,P_0}{P_0 + (L-P_0)\, e^{-\,L\,k\,(t-t_0)}} \\
&= \frac{L}{1 + (\frac{L}{P_0} - 1)\, e^{-\,L\,k\,(t-t_0)}}.
\end{aligned} \tag{4.8}$$

Now let's use (4.8) to check our sketches given in Fig. 4.2. Since both L and k are positive, thus $e^{-\,L\,k\,(t-t_0)}$ decays with t ($t > t_0$). Hence, for large t, $P \approx L$ since $e^{-\,L\,k\,(t-t_0)} \approx 0$ irrespective of P_0 (see Fig. 4.2 (b)). Moreover, we also see from (4.8) that if $P_0 = L$, then

$P = L$. Conversely, if $P = L$, then $P_0 = L$. This means that no solution other than $P = L$ ever actually touches $P = L$, implying that no curve crosses the line $P = L$ (Fig. 4.2 (a)). Therefore, $P = L$ is an attractor. (An infinite time to reach.)

4.2 Logistic Map

If we want a numerical solution to the logistic equation (4.3), computers only use floating point arithmetic (or rational truncations). That means no computer "knows" what an irrational number is. We work in practice with real computations with rational approximations to real numbers. In fact, we work with a finite subset of rational numbers. This implies that we do not, in a computer, solve equations or differential equations exactly. That is, one uses what they need at time t to estimate what the system will do at time $t + \Delta t$, for small Δt. For example, for the simple case of

$$\frac{dP}{dt} = k\,P$$

We can estimate $P(t+\Delta t)$ by $\tilde{k}\,P(t)$ where $\tilde{k}$ is to be determined. By using Taylor series, we obtain

$$\begin{aligned} P(t+\Delta t) &= P(t) + \left.\frac{dP}{dt}\right|_t \Delta t + O((\Delta t)^2) \\ &= P(t)\ (1 + k\,\Delta t) + O((\Delta t)^2) \\ &\approx \tilde{k}\,P(t). \end{aligned} \tag{4.9}$$

Thus, we have $P(t+\Delta t) = \tilde{k}\,P(t)$ or

$$P(t + n\,\Delta t) = \tilde{k}\,P\,(t + (n-1)\Delta t) \qquad \text{or simply} \quad P_n = \tilde{k}\,P_{n-1} \tag{4.10}$$

which can be considered as a *discrete map*. Then

$$\begin{aligned} P(t+2\,\Delta t) &= \tilde{k}\,P(t+\Delta t) = \tilde{k}^2 P(t) \\ P(t+3\,\Delta t) &= \tilde{k}^3 P(t) \\ &\vdots \\ P(t+n\,\Delta t) &= \tilde{k}^n P(t) \end{aligned}$$

or simply we can write

$$P_n = \tilde{k}^n \, P_0. \tag{4.11}$$

Note: This has the same exponential behavior as the continuous picture $P = P_0 \, e^{k\,t}$.

There are, of course, other numerical schemes, but this one is perhaps the simplest and more "real".

Since population is, after all, not a continuous function; we "move" from generation to generation discretely. The continuous form is in some sense an "approximation" of the discrete map, which is the real approximation then? We shall leave this question unanswered, since continuous and discrete systems are both just simulations after all in this case.

What is "map"? By a map, we mean either a (continuous) function $f(x)$ or the difference equation $x_{n+1} = f(x_n)$.

Now let's look for a discrete map for the logistic equation with the same procedure as above:

$$\begin{aligned} P(t+\Delta t) &= P(t) + \left.\frac{dP}{dt}\right|_t \Delta t + O((\Delta t)^2) \\ &= P(t) + k\,P(t)\,(L - P(t))\,\Delta t + O((\Delta t)^2) \\ &= (1 + L\,k\,\Delta t)\,P(t) - k\,\Delta t\,(P(t))^2 + O((\Delta t)^2) \\ &= P(t)\,[\,(1 + L\,k\,\Delta t) - k\,\Delta t\,P(t)\,] + O((\Delta t)^2) \\ &= (1 + L\,k\,\Delta t)\,P(t)\left[1 - \frac{k\,\Delta t}{1 + L\,k\,\Delta t}\,P(t)\right] + O((\Delta t)^2). \end{aligned} \tag{4.12}$$

Let

$$\mu = (1 + L\,k\,\Delta t) \qquad \text{and} \qquad \tilde{\mu} = \frac{k\,\Delta t}{\mu}, \tag{4.13}$$

then from (4.12) we have a discrete map

$$P_{n+1} = \mu\,P_n\,(1 - \tilde{\mu}\,P_n). \tag{4.14}$$

For simplicity, let

$$x_n = \tilde{\mu}\,P_n \qquad \text{and so} \qquad x_{n+1} = \tilde{\mu}\,P_{n+1}\,, \tag{4.15}$$

then equation (4.14) becomes

$$x_{n+1} = \mu\,x_n\,(1 - x_n) \tag{4.16}$$

which is called *logistic map*. It is also known as *quadratic map*. For convenience, we use $f_\mu(x_n)$ to denote the function on the right-hand-side (RHS) of equation (4.16). This function $f_\mu(x_n)$ (map) is just one single hump (Fig. 4.3).

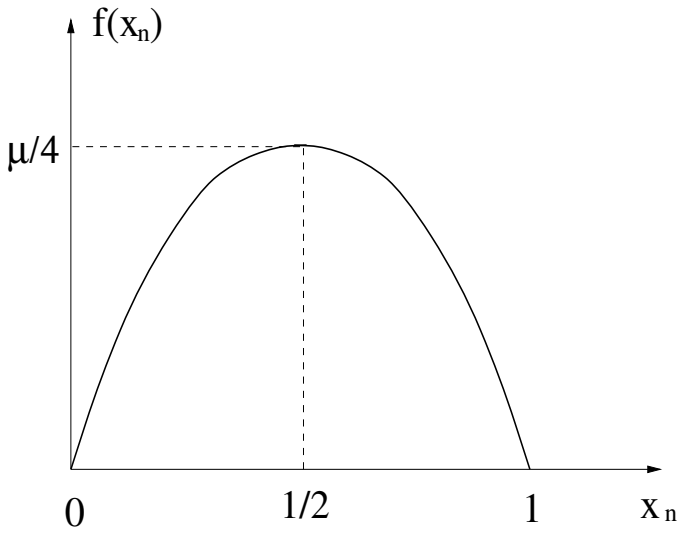


Fig. 4.3 The logistic map.

This discrete map, it turns out, has some extraordinary behavior that differs from its corresponding differential equation. Although, especially, for small μ, that is much in common, as must be the case.

As we have seen, a discrete map generates a sequence as a function of n ($n = 0,\ 1,\ 2,\ \cdots$). In general, we take a value x_n and apply a function $f(\bullet)$ (this is so you don't get hang up on what the independent variable is) to it to get x_{n+1}, then reapply this function to x_{n+1} and so on. Thus,

$$x_{n+1} = f(x_n) = f\left(f(x_{n-1})\right) = f\left(\left(f(x_{n-2})\right)\right) = \cdots$$

and we have the sequence

$$\begin{aligned} x_1 &= f(x_0) = \mu\, x_0\,(1 - x_0) = f^1(x_0) \\ x_2 &= f(x_1) = \mu\, x_1\,(1 - x_1) = f^2(x_0) \\ &\vdots \\ x_n &= f(x_n) = \mu\, x_n\,(1 - x_n) = f^n(x_0). \end{aligned} \tag{4.17}$$

We have seen fixed points defined in continuous logistic system $\dfrac{dP}{dt} = k\,P\,(L - P)$ as $P = 0$ or $P = L$. Similarly, we can also define fixed points for a discrete map. Discrete maps have fixed points when there is no change, just like with differential equations. The **fixed point** of a map, x^*, is defined as a point satisfying

$$f(x^*) = x^*. \tag{4.18}$$

So if $x_n = x^*$ then $x_{n+1} = f(x_n) = f(x^*) = x^*$; hence the orbit remains at x^* for all future times. For the logistic map: $x_{n+1} = \mu\, x_n\,(1 - x_n)$, clearly, $x_n = 0$ is a fixed point. This leads to $P_n = 0$ (since $x_n = \tilde{\mu}\, P_n$), corresponding to $P = 0$ which is the *same as the differential equation.* There is another fixed point which one must find by finding x such that

$$x = f(x) \qquad \text{Note : the subscript is irrelevant}$$

or

$$x = \mu\; x\,(1 - x). \tag{4.19}$$

So we have $x = 0$. This is the one $x_n = 0$ we just found; or

$$1 - \mu\,(1 - x) = 0 \qquad \text{i.e., } x = 1 - \frac{1}{\mu}. \tag{4.20}$$

Recall that $\mu = 1 + L\,k\,\Delta t$ and $\tilde{\mu} = \dfrac{k\,\Delta t}{\mu}$, then (4.20) implies that

$$x = \frac{\mu - 1}{\mu} = \frac{L\,k\,\Delta t}{1 + L\,k\,\Delta t} = \tilde{\mu}\,L. \tag{4.21}$$

On the other hand, from $x_n = \tilde{\mu}\, P_n$, we have

$$x = \tilde{\mu}\, P. \tag{4.22}$$

Comparing (4.22) with (4.21) yields $P = L$. Hence, this non-trivial fixed point given by (4.20) corresponds to $P = L$, the fixed point of the original differential equation.

But are these fixed points stable or not? Does this *stability* or *instability* correspond to the differential equation? How do we determine this? A Taylor series is a simple way to derive the stability of a fixed point. (There are certain other approaches, such as using mean value theorem.)

Suppose we have a fixed point x^* satisfying the discrete map $x^* = f(x^*)$. To determine the stability of x^*, consider a nearby point $x_n = x^* + \eta_n$ and we ask whether the point is *attracted to* or *repelled from* x^*. In other words, does the derivation of η_n *grow* or *decay* as n increases? Substitute $x_{n+1} = x^* + \eta_{n+1}$ into the map $x_{n+1} = f(x_n)$ to obtain

$$x^* + \eta_{n+1} = x_{n+1} = f(x^* + \eta_n) = f(x^*) + f'(x^*)\,\eta_n + O(\eta_n^2)$$

which is reduced to

$$\eta_{n+1} = f'(x^*)\,\eta_n + O(\eta_n^2) \tag{4.23}$$

due to $x^* = f(x^*)$.

Suppose we can choose η_n small enough (i.e., after enough long transient period) so that we may neglect the $O(\eta_n^2)$ terms. then we have

$$\eta_{n+1} = f'(x^*)\,\eta_n \tag{4.24}$$

which is called *linearized map*. The $f'(x^*)$ is called its *multiplier* (we don't call it as *eigenvalue* which we want to be reserved for the continuous systems), usually denoted as λ. The solution of the linearized map can be easily found as

$$\begin{aligned} \eta_1 &= \lambda\,\eta_0 \\ \eta_2 &= \lambda\,\eta_1 = \lambda^2\,\eta_0 \\ &\vdots \\ \eta_n &= \lambda^n\,\eta_0. \end{aligned}$$

Thus, if $|\lambda| = |f'(x^*)| < 1$, then $\lim_{n\to\infty} \eta_n = 0$, and the fixed point x^* is **linearly stable**. Conversely, if $|\lambda| = |f'(x^*)| > 1$, then $\lim_{n\to\infty} \eta_n \to \infty$, so the fixed point is unstable. Although these conclusions about *local stability* are based on linearization, they can be proven to hold for the original nonlinear map. But the linearization tells us nothing about the **marginal** case $|f'(x^*)| = 1$; then the neglected $O(\eta_n^2)$ terms determine the local stability.

Now let's use the above results to consider the stability of the fixed points of the logistic map:

$$x_{n+1} = \mu\,x_n\,(1 - x_n).$$

We have found the fixed points of the logistic map as

$$x^* = 0 \qquad \text{and} \qquad x^* = 1 - \frac{1}{\mu}. \tag{4.25}$$

Here,

$$f(x) = \mu\, x\,(1 - x) \qquad \text{so} \quad f\,'(x) = \mu - 2\,\mu\, x. \tag{4.26}$$

For $x^* = 0$, $f\,'(0) = \mu$, so the fixed point $x^* = 0$ is unstable for $|\mu| > 1$. This agrees with the original DE that $P = 0$ is unstable since $\mu > 1 \Longleftrightarrow k > 0$.

Next, for the fixed point $x^* = 1 - \dfrac{1}{\mu}$, $f\,'\left(1 - \dfrac{1}{\mu}\right) = \mu - 2\,\mu\left(1 - \dfrac{1}{\mu}\right) = 2 - \mu$. If we want $\left|f\,'\left(1 - \dfrac{1}{\mu}\right)\right| < 1$, then $|2 - \mu| < 1$ or

$$-1 < \mu - 2 < 1 \qquad \Longrightarrow \qquad 1 < \mu < 3. \tag{4.27}$$

Hence,

$$\text{Fixed point } x^* = 1 - \frac{1}{\mu} \begin{cases} \text{stable} & \text{if} \quad 1 < \mu < 3; \\ \text{unstable} & \text{if} \quad \mu < 1 \ \text{ or } \ \mu > 3. \end{cases} \tag{4.28}$$

when $\mu = 1$ $x^* = 0$ and $f\,'(0) = 1$, so it is a marginal case; when $\mu = 3$ $x^* = \dfrac{2}{3}$ and $f\,'\left(\dfrac{2}{3}\right) = -1$, so it is also a marginal case.

For the population problem or the logistic map, we consider $0 \le x_n \le 1$ and suppose that the intrinsic growth rate $\mu \ge 0$. As shown in Fig. 4.3, the graph of (4.16) is a parabola with a maximum value of $\dfrac{\mu}{4}$ at $x = \dfrac{1}{2}$. So we restrict the control parameter μ to the range $0 \le \mu \le 4$ so that (4.16) maps the interval $0 \le x \le 1$ into itself. The behavior is much less interesting for other values of x and μ. (It can be shown that for any $\mu > 0$, if $x_n > 1$ for some n, then subsequent iterations diverge toward $-\infty$. For the application of population problem, this means the population goes extinct.)

Fixed points with multiplier $\lambda = 0$ are called **superstable** because the convergence to the fixed point is *exponentially* rapid. It can be proved that in the superstable case, the perturbations decay like $\bar{x}_n \sim \eta_0^{2^n}$, which is much faster than the usual $\eta_n \sim \lambda^n\, \eta_0$ at an ordinary stable point. The proof is given below: since the Taylor series of x_{n+1} near x^* yields

$$\eta_{n+1} = \frac{1}{2}\, f\,''(x^*)\, \eta_n^2 + O(\eta_n^3)\,, \qquad (\text{due to } f\,'(x^*) = 0) \tag{4.29}$$

so, we have the following approximate formulae:

$$\begin{aligned}
\eta_1 &= \frac{1}{2} f''(x^*)\, \eta_0^2 \\
\eta_2 &= \frac{1}{2} f''(x^*)\, \eta_1^2 = \left(\frac{1}{2} f''(x^*)\right)^3 \eta_0^4 \\
&\cdots \\
\eta_n &= \left(\frac{1}{2} f''(x^*)\right)^{2^n-1} \eta_0^{2^n} = \frac{2}{f''(x^*)} \left(\frac{1}{2} f''(x^*)\, \eta_0\right)^{2^n}
\end{aligned} \tag{4.30}$$

which implies that η_n decays like $\eta_0^{2^n}$ as long as we choose η_0 such that $|f''(x^*)\, \eta_0| < 2$.

Before we further study the logistic map, we now introduce a graphical method which is very useful in considering a discrete map.

4.3 Plotting Graph of Iterates (Cobweb)

Cobweb is a graphical method which is usually applied in the analysis of maps. Given $x_{n+1} = f(x_n)$ and an initial condition x_0, draw a vertical line until it intersects the graph of f; that height is the output x_1 (see Fig. 4.4). At this stage we could return to the horizontal axis and repeat the procedure to get x_2 from x_1, but it is more convenient simply to trace a horizontal line till it intersects the diagonal line $x_{n+1} = x_n$, and then move vertically to the curve again. Repeat the process n times to generate the first n points in the orbit.

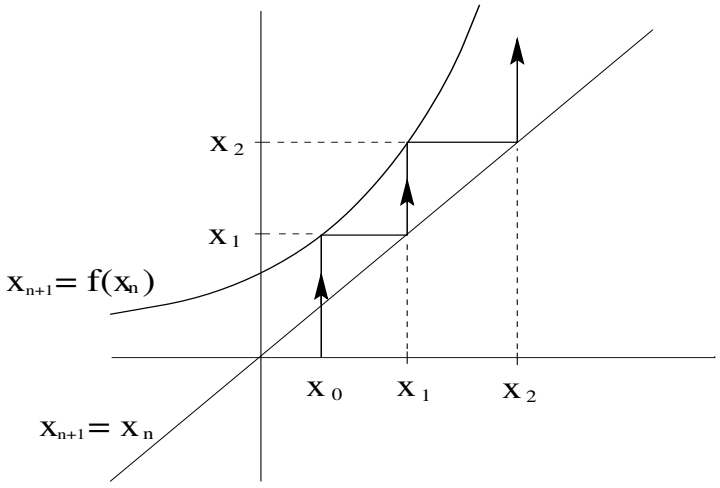


Fig. 4.4 The cobweb for a mapping $x_{n+1} = f(x_n)$.

Cobwebs are useful because they allow us to see global behavior at a glance, thereby supplementing the local information available from the linearization. Cobwebs become even more valuable when linear analysis fails, as in the next example.

Consider the map

$$x_{n+1} = \sin x_n \,. \tag{4.31}$$

First note that $x^* = 0$ is a unique fixed point. Next, we shall show that the stability of the fixed point is not determined by the linearization and use a cobweb to show that $x^* = 0$ is stable – in fact, *globally* stable.

The multiplier at $x^* = 0$ is $f'(0) = \cos(0) = 1$, which is a marginal case where linear analysis is inconclusive. However, the cobweb of Fig. 4.5 shows that $x^* = 0$ is locally stable; the orbit slowly rattles down the narrow channel, and heads monotonically for the fixed point. (A similar picture is obtained for $x_0 < 0$.)

To see that the stability is global, we have to show that *all* orbits satisfy $x_n \to 0$. But for any x_0, the first iterate is sent immediately to the interval $-1 \le x_1 \le 1$ since $|\sin x| \le 1$. The cobweb in that interval looks qualitatively like Fig. 4.5, so convergence is assured.

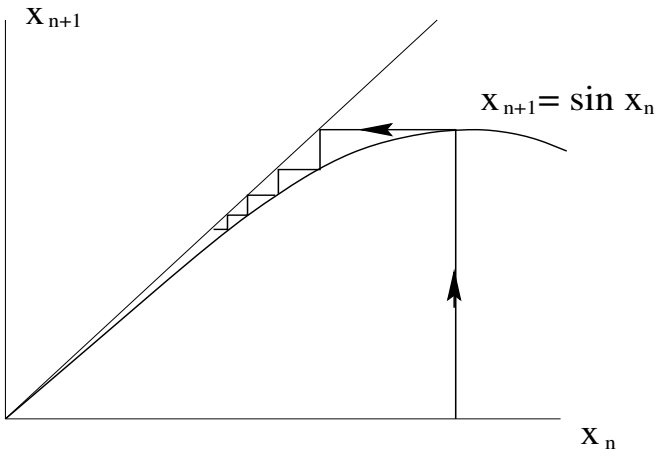


Fig. 4.5 The cobweb for the map $x_{n+1} = \sin x_n$.

Now let's consider the logistic map

$$x_{n+1} \;=\; \mu\, x_n\,(1 - x_n)$$

$$
\begin{aligned}
&= -\mu\left(x_n^2 - x_n + \frac{1}{4} - \frac{1}{4}\right) \\
&= -\mu\left(x_n^2 - \frac{1}{2}\right)^2 + \frac{1}{4}\mu \\
&\equiv f(x_n) \qquad (4.32)
\end{aligned}
$$

which is a parabola with its vertex at $(\frac{1}{2}, \frac{1}{4}\mu)$. Fig. 4.6 show the logistic map $f(x_n)$ for different values of μ.

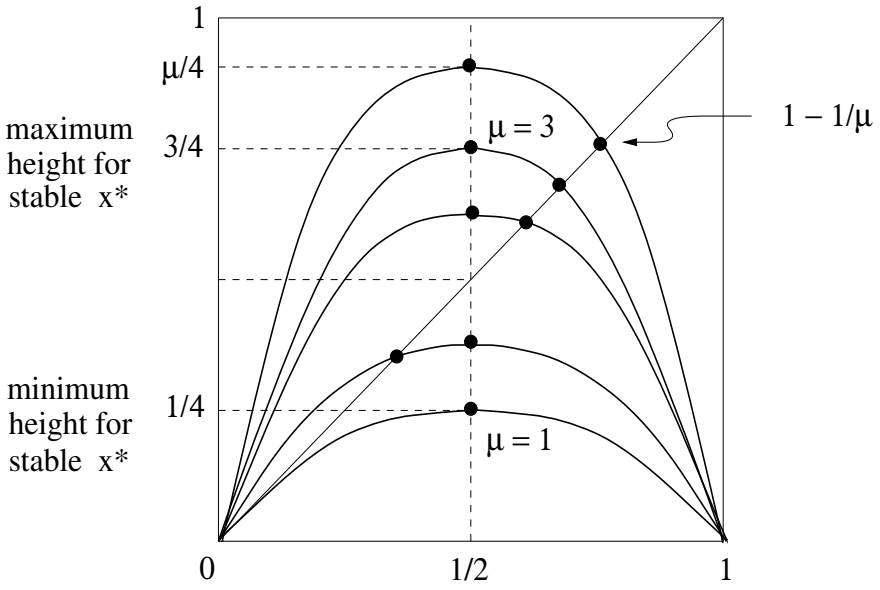


Fig. 4.6 The logistic map for different values of μ.

Recall that the fixed point $x^* = 1 - \frac{1}{\mu}$ is stable when $1 < \mu < 3$.

Notes:

1) This map needs to be redefined for $\mu > 4$ because values of $x > 1$ can occur which ultimately divergent.

2) Slope outside $0 \le x \le 1$ has magnitude strictly greater than 1.

When $1 < \mu < 3$, all initial conditions, except for $x_0 = 0$ and $x_0 = 1$, approach the fixed point $x^* = 1 - \frac{1}{\mu}$. See Fig. 4.7. It should be noted that $x_0 = 1$ goes to the fixed point $x^* = 0$ immediately.

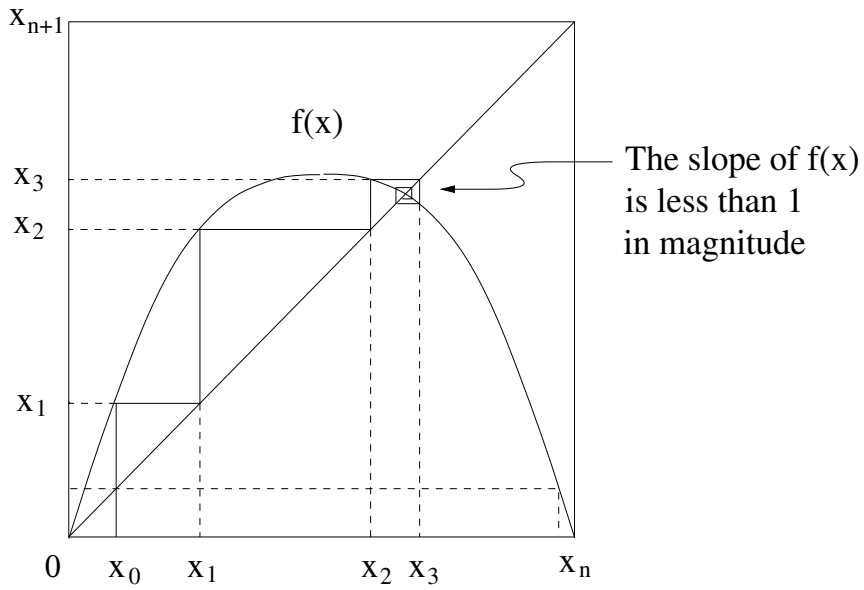


Fig. 4.7 Stable fixed point $x^* = 1 - \dfrac{1}{\mu}$.

Question: *What happens to the population as* $n \to \infty$, *and as* μ *varies?*

Let's see what happens when μ is varied.

Case 1: For small growth rate $\mu < 1$ (say $\mu = \dfrac{1}{2}$), the population always dies out no matter what the initial condition is; i.e., $x_n \to 0$ as $n \to \infty$. See Fig. 4.8 for the case $\mu = \dfrac{1}{2}$, $x_0 = \dfrac{1}{2}$.

Case 2: For $1 < \mu < 3$ (say $\mu = 2$), the population grows and eventually reaches a nonzero steady state. In the limit, all initial conditions (all x_0) lead to x^*. Even though it may take an infinite number of iterates. We say this is the "ω" limit set of the map. (Similarly, we may define so called "α" limit set.) The case $x_0 = \dfrac{1}{3}$ is shown in Fig. 4.9. Again, no sign of anything is unusual. It should be noted that since $|f'(x^*)| < 1$ when $1 < \mu < 3$, so the fixed point x^* is stable. Moreover, let $\eta_n = x_n - x^*$, we have

$$\eta_{n+1} = f'(x^*)\,\eta_n.$$

Since $f'(x^*) > 0$ when $1 < \mu \leq 2$ and $f'(x^*) < 0$ when $2 < \mu < 3$, so η_{n+1} and η_n have the same sign for $1 < \mu \leq 2$ (see Fig. 4.9 in which $\mu = 2$); but have opposite sign for $2 < \mu < 3$ as shown in Fig. 4.10 where $\mu = 2.8$.

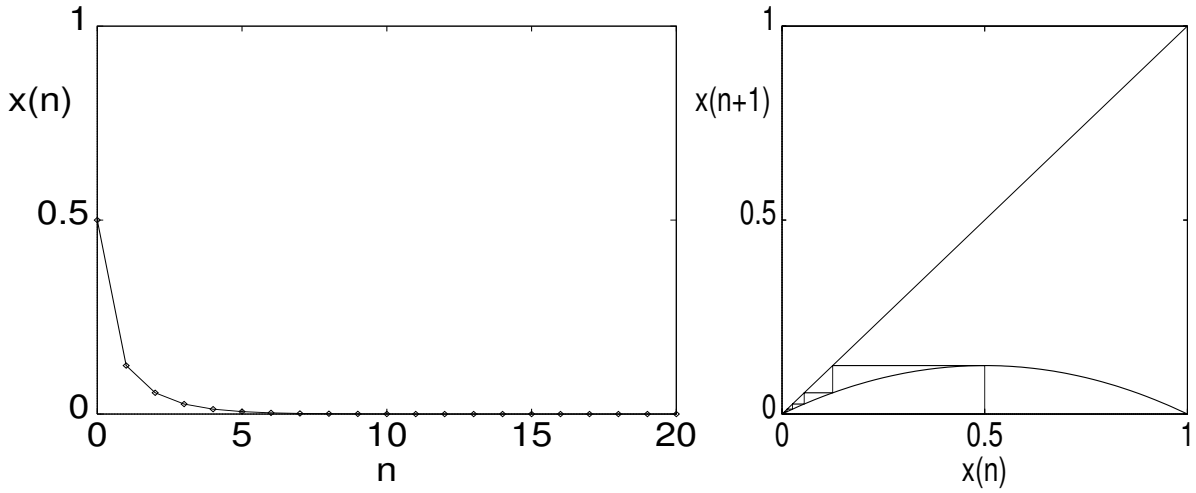


Fig. 4.8 The time series of the logistic map when $\mu = \frac{1}{2}$, $x_0 = \frac{1}{2}$.

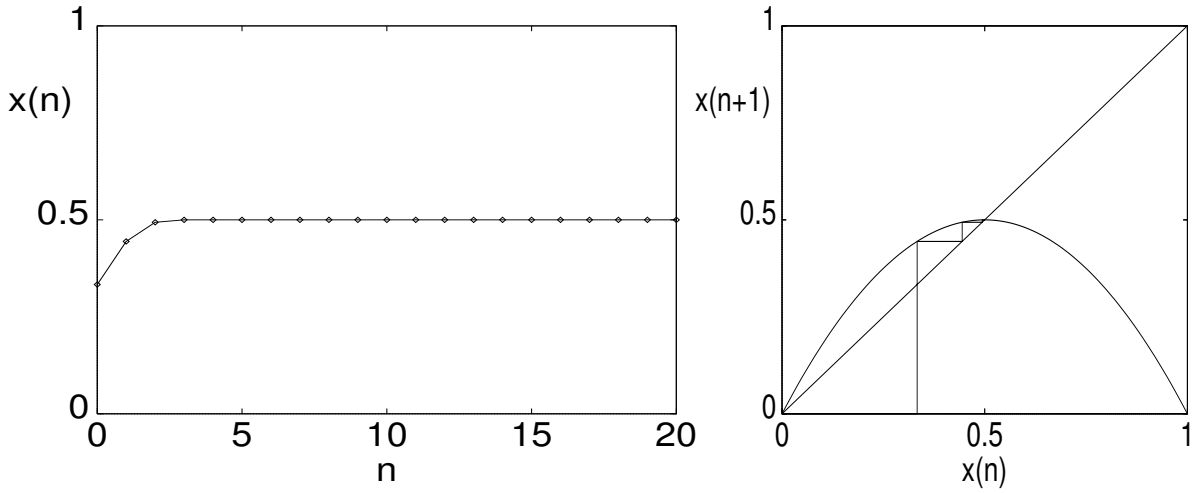


Fig. 4.9 The time series of the logistic map when $\mu = 2$, $x_0 = \frac{1}{3}$.

From the above two cases, it can be seen that there exist two fixed points: $x^* = 0$ and $x^* = 1 - \frac{1}{\mu}$, which are determined from the equation $x^* = \mu\, x^*(1 - x^*)$. This is similar to the continuous logistic system (4.3) if one notices from equation (4.13) that $\mu < 1 \iff k < 0$ and $\mu > 1 \iff k > 0$. However, note that initial condition may "cross" $x = x^*$, see Fig. 4.10. Also note that as we move backwards there are always 2 possible x_n's (see the cobweb in Fig. 4.7) from which one can arrive at any x_n. We say that the logistic map is

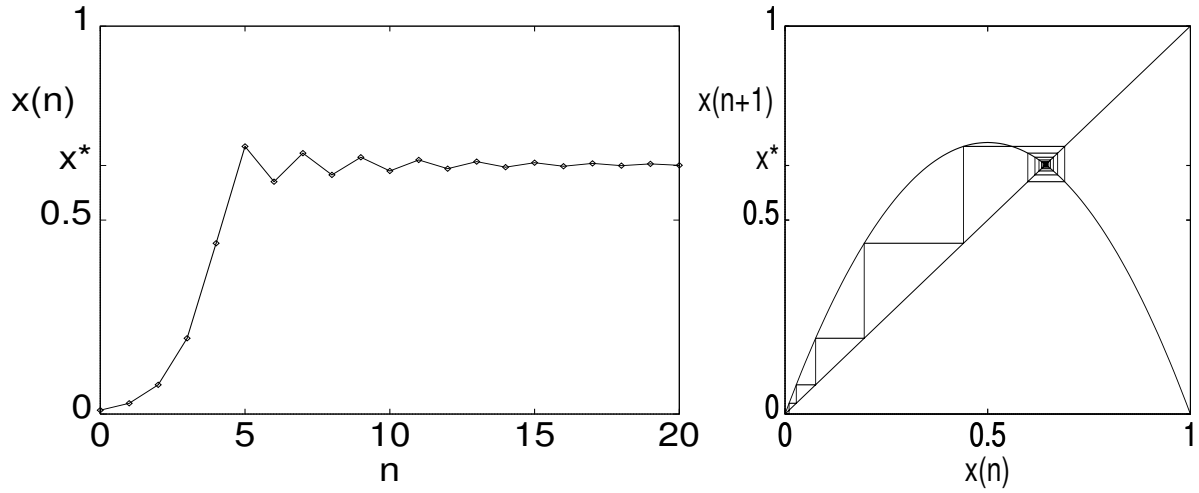


Fig. 4.10 The time series of the logistic map when $\mu = 2.8\,,\ x_0 = \dfrac{1}{3}$.

non-invertible. The logistic equation is invertible by virtue of uniqueness.

4.4 Bifurcation: 2-Cycle

As μ increases, the fixed point $x^* = 1 - \dfrac{1}{\mu}$ becomes unstable for $\mu > 3$.

Case 3: For $3 < \mu < \mu_c$ ($\mu_c = 3.569946\cdots$), the system exhibits some interesting phenomena which cannot be observed from the continuous logistic system. For example, when $\mu = 3.3$, the population builds up again but now *oscillates* about the nonzero steady state, alternating between a large population in one generation and a smaller population in the next (Fig. 4.11). This type of oscillation, in which x_n repeats every *two* iterations, is called a cycle. In particular it is known as a **period-2 cycle**. (Note that for a continuous system to have periodic solutions, its dimension of phase space must be at least 2.)

In terms of our "phase space", all initial conditions (except for the unstable fixed points $x = 0,\ x = 1 - \dfrac{1}{\mu}$ as well as $x = 1$) approaches this 2-cycle (Fig. 4.12).

The "ω" limit set is composed of 2 values, in this case, x_+ and x_-. The *transition* as μ increases from 1 to 2 values is known as a *bifurcation*. The value of μ where this transition occurs is called as a *bifurcation point*. See Fig. 4.13.

The bifurcation represents a qualitative change in the nature of the behavior of the

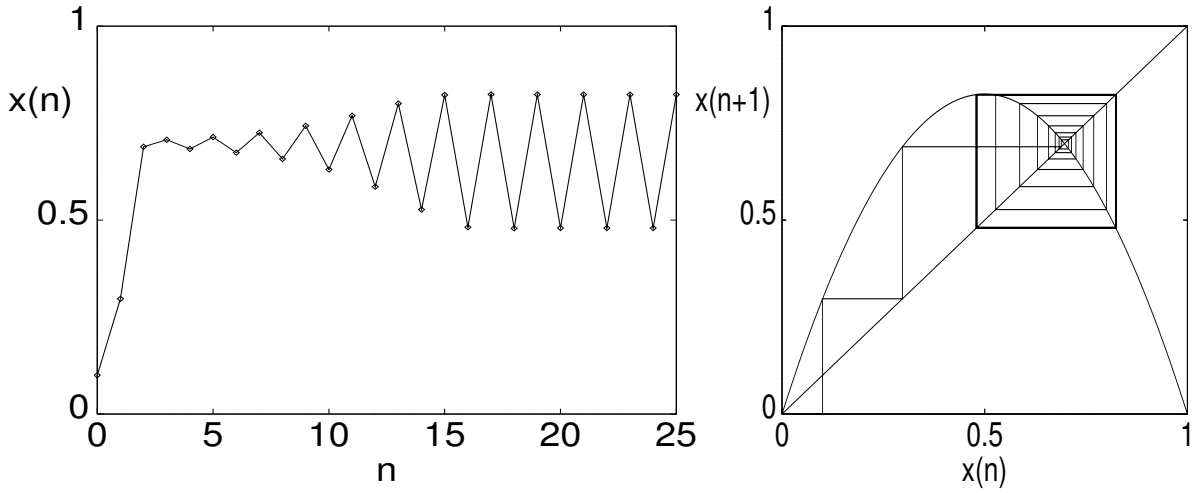


Fig. 4.11 Bifurcation of period-2 cycle when $\mu = 3.3$.

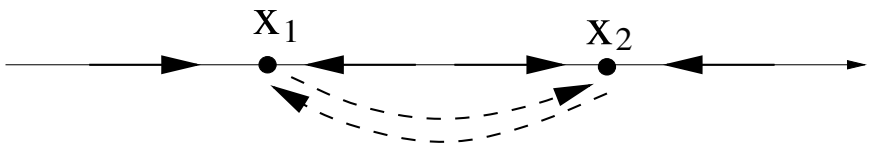


Fig. 4.12 The phase space of 2-cycle.

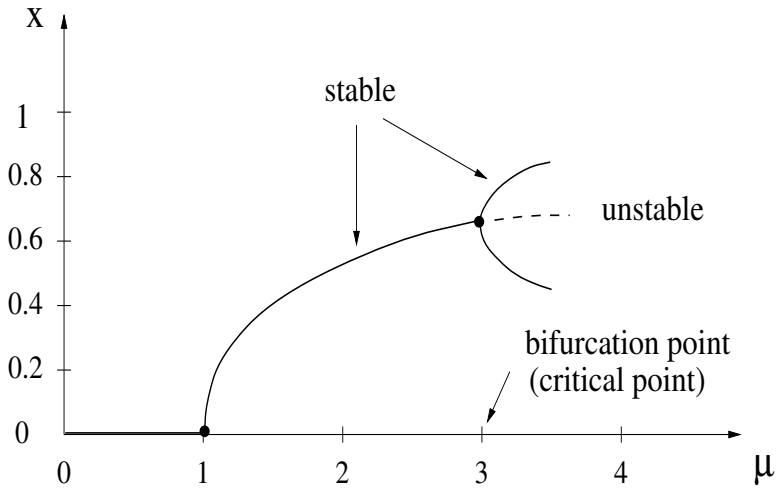


Fig. 4.13 Bifurcation to 2-cycle.

map: an asymptotic relaxation becomes a relaxation to a steady oscillation. This behavior is foreign to the DE from which we started. However, we'll see later that DEs can also

exhibit rich bifurcation and instability phenomena.

In this case, the values that become the limit set are not fixed points of $x_{n+1} = f(x_n)$, however, given this periodic nature, they *are* fixed points of

$$\tilde{x}_{n+1} = f(f(\tilde{x}_n)) \equiv f^2(\tilde{x}_n) \tag{4.33}$$

where $\tilde{x}_{n+1}$ are the odd or even iterates of the sequence $x_0,\ x_1, \cdots$; that is,

$$\begin{aligned} x_+ &= \mu\, x_- \,(1 - x_-) \\ x_- &= \mu\, x_+ \,(1 - x_+) \end{aligned} \tag{4.34}$$

which clearly indicates that $x_+ = f^2(x_+)$ and $x_- = f^2(x_-)$. Hence, x_+ and x_- are fixed points of the map $f^2(x)$, i.e., they are 2-cycle of $f(x)$. It should be noted that $x = 0$ and $x = 1 - \dfrac{1}{\mu}$ are also fixed points of $f^2(x)$ but they are unstable. The map $f^2(x)$ is shown in Fig. 4.14.

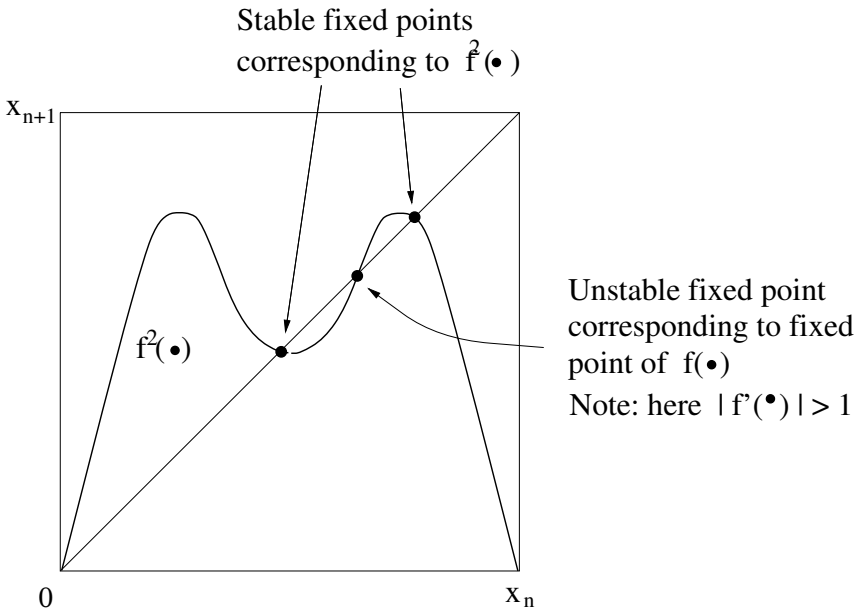


Fig. 4.14 Map $f^2(x)$.

Now we show that the fixed points (2-cycle) x_+ and x_- are stable when $3 < \mu < 1+\sqrt{6}$. Consider the logistic map

$$x_{n+1} = \mu\, x_n\,(1 - x_n) = f(x_n).$$

To find 2-cycle, let

$$p = f(q) = \mu\, q\,(1-q) \qquad \text{and} \qquad q = f(p) = \mu\, p\,(1-p). \tag{4.35}$$

Then subtracting one of the above equations from the other yields

$$p - q = \mu\,[\,q\,(1-q) - p\,(1-p)\,] = \mu\,(q-p)\,(1-q-p)$$

$$\Longrightarrow$$

$$(p-q)\,[\,1 + \mu\,(1-p-q)\,] = 0 \tag{4.36}$$

which, in turn, results in

$$q = p \qquad \text{or} \qquad q = 1 + \frac{1}{\mu} - p. \tag{4.37}$$

Since $q = p$ is not 2-cycle, so we only take the solution $q = 1 + \dfrac{1}{\mu} - p$ and substitute it into the first equation of (4.35) to obtain

$$\mu\, p^2 - (1+\mu)\, p + 1 + \frac{1}{\mu} = 0. \tag{4.38}$$

Thus,

$$p_{\pm} = \frac{\mu + 1 \pm \sqrt{(\mu+1)\,(\mu-3)}}{2\,\mu} \tag{4.39}$$

which are real for $\mu > 3$. q can then be found as $q_{\pm} = p_{\mp}$. So the fixed points are $p_{\pm}$. To find the stability of $p_{\pm}$, we need to compute $(f^2(p_{\pm}))'$ which is equal to

$$(f^2(p_{\pm}))' = \frac{d}{dx}\,(f(f(x)))\,|_{x=p_{\pm}} = f'(f(p_{\pm}))\, f'(p_{\pm}) = f'(q_{\pm})\, f'(p_{\pm}). \tag{4.40}$$

After carrying out the differentiations and substituting for $p_{\pm}$ and $q_{\pm}$ (remember $q_{\pm} = p_{\mp}$), we obtain

$$\begin{aligned}
(f^2(p_{\pm}))' &= f(f(p_{\pm}))' = f'(f(p_{\pm}))\, f'(p_{\pm}) = f'(q_{\pm})\, f'(p_{\pm}) \\
&= \mu\,(1 - 2\,q_{\pm})\,\mu\,(1 - 2\,p_{\pm}) \\
&= \mu^2\,[\,1 - 2\,(p_{\pm} + q_{\pm}) + 4\,p_{\pm}\,q_{\pm}\,] \\
&= \mu^2\,[\,1 - 2\,(p_{\pm} + p_{\mp}) + 4\,p_{\pm}\,p_{\mp}\,] \\
&= \mu^2\left[1 - 2\left(\frac{\mu+1}{\mu}\right) + 4\left(\frac{\mu+1}{\mu^2}\right)\right] \\
&= 4 + 2\,\mu - \mu^2.
\end{aligned} \tag{4.41}$$

Hence, the 2-cycle is linearly stable for $|4+2\,\mu-\mu^2|<1$; i.e., $-1<4+2\,\mu-\mu^2<1$. Form the first inequality we have

$$\mu^2-2\,\mu-5<0 \qquad \Longrightarrow \qquad 1-\sqrt{6}<\mu<1+\sqrt{6};$$

while from the second inequality we found

$$\mu^2-2\,\mu-3>0 \qquad \Longrightarrow \qquad \mu<1 \quad \text{or} \quad \mu>3.$$

Combining these above inequalities yields the stability condition for the 2-cycle:

$$3<\mu<1+\sqrt{6}. \tag{4.42}$$

4.5 Period-Doubling: Bifurcation to Chaos

For $\mu>1+\sqrt{6}$, x_+ and x_- become unstable fixed points of $f^2(x)$. The 2-cycle disappears and is replaced by a 4-cycle: another bifurcation (Fig. 4.15).

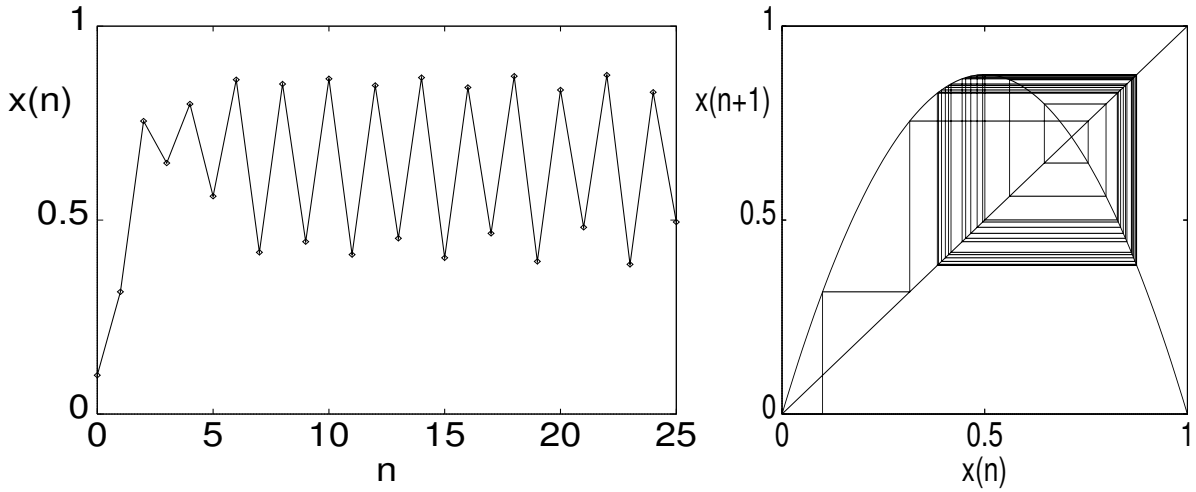


Fig. 4.15 Bifurcation of period-4 cycle when $\mu=3.5$.

These solutions are stable fixed points of the map

$$\tilde{x}_{n+1}=f^2(f^2(\tilde{x}_n))=f^4(\tilde{x}_n). \tag{4.43}$$

The other fixed points 0, $1-\dfrac{1}{\mu}$, x_+ and x_- are also the fixed points of the map $f^4(x)$ but they are *unstable*. The new bifurcating 4 fixed points are stable: a 4-cycle for the map $f(x)$.

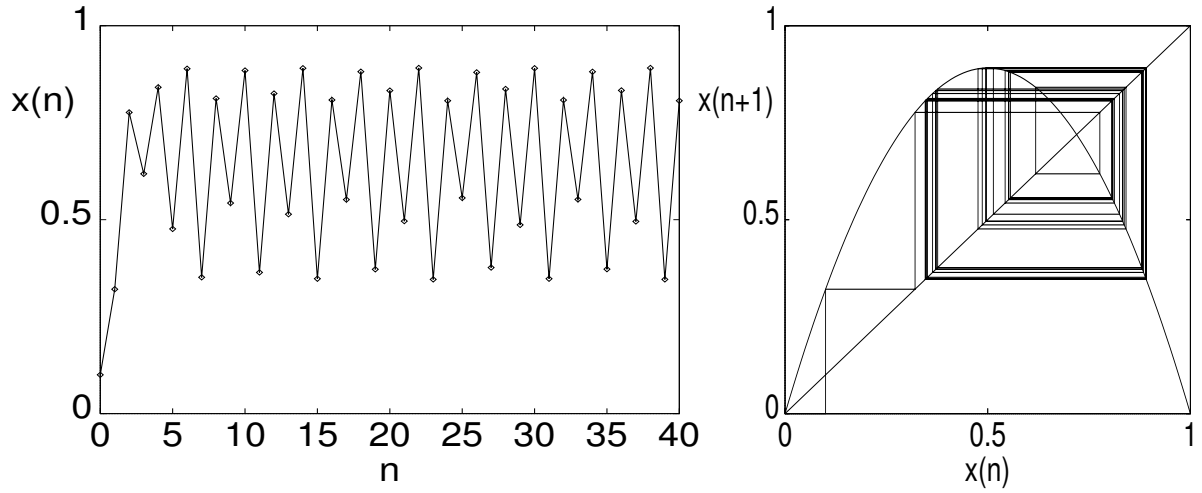


Fig. 4.16 Bifurcation of period-16 cycle when $\mu = 3.5644$.

However they are only stable for $1+\sqrt{6} = 2.44949\cdots < \mu < 3.544090\cdots$. There is, in fact, another bifurcation at $\mu = 3.544090\cdots$. The iterates transform into an 8-cycle. It seems that at every bifurcation point: the cycle is doubled (or period doubled). Indeed, further **period-doubling** to cycles 16 (Fig. 4.16), 32, etc. occurs as μ increases. Specifically, let μ_n denotes the value of μ where a 2^n-cycle first appears. Then computer experiments reveal that

$\mu_2 = 3$	(period 2 is born)
$\mu_4 = 3.449\cdots$	4
$\mu_8 = 3.54409\cdots$	8
$\mu_{16} = 3.5644\cdots$	16
$\mu_{32} = 3.568759\cdots$	32
$\mu_{64} = 3.569692\cdots$	32
$\vdots$	$\vdots$

It should be clear that as n goes large these values are approaching a limit

$$\mu_\infty = \lim_{n\to\infty} \mu_n = 3.569946\cdots \tag{4.44}$$

This is just a number, but it has important implications! It means that we have an "∞"-cycle for that value of μ. Note that the successive bifurcations come faster and faster. The

convergence is essentially geometric: in the limit of large n, the distance between successive transitions shrinks by a constant factor

$$\delta = \lim_{n\to\infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n} = 4.6692016091\cdots \tag{4.45}$$

In fact, the same convergence rate (number) appears *no matter what* **one-hump maps** or **Unimodal maps** *are iterated*! In this sense, this number is **universal** and called *Feigenbaum constant*. It is a new mathematical constant, as basic to period-doubling as π is to circles.

The above period-doubling bifurcations can be illustrated in the state-parameter space (Fig. 4.17).

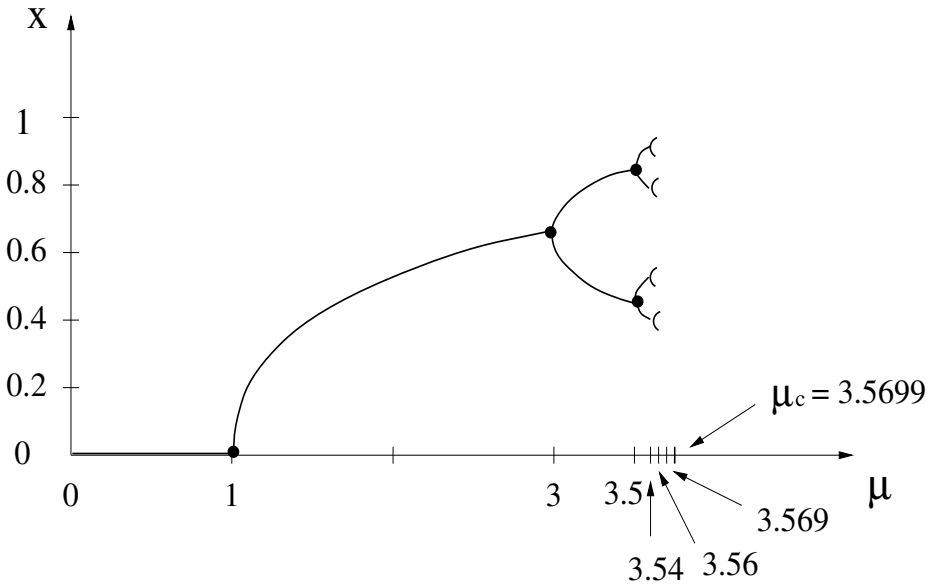


Fig. 4.17 Period-doubling bifurcations.

It should be noted here that μ_∞ is less than 4, which means that all iterates are confined and there are infinite points in the ω limit set. When $\mu > \mu_\infty$, the behavior turns out to be chaos. We note that the ω limit set before μ_∞ is zero dimensional – just a bunch of separate finite points. But at μ_∞ the ω set is something between a line and a bunch of separate points. It has a dimension approximately 0.53. (Not an integer! More on non-integer dimension later.)

The process to an "∞" cycle is called a **period-doubling cascade**.

4.6 "Fingerprint" of Chaos: Failure of Naive Determinism.

From the qualitative point of view, $\mu = 4$ (which is the highest value of μ that is meaningful) is well into this "peculiar" chaotic domain. Now the equation (4.16) becomes

$$x_{n+1} = 4\,x_n\,(1 - x_n). \tag{4.46}$$

By using the transformation

$$x_n = \sin^2\theta_n, \tag{4.47}$$

one can transform equation (4.46) into

$$\begin{aligned} \sin^2\theta_{n+1} &= 4\,\sin^2\theta_n\left(1 - \sin^2\theta_n\right) \\ &= 4\,\sin^2\theta_n\,\cos^2\theta_n \\ &= (2\,\sin\theta_n\,\cos\theta_n)^2 \\ &= (\sin 2\theta_n)^2 \end{aligned} \tag{4.48}$$

or

$$\sin(\theta_{n+1}) = \pm\,\sin(2\theta_n) \tag{4.49}$$

i.e.,

$$\theta_{n+1} = \pm\,2\,\theta_n + k\,2\pi \qquad k = 0,\,\pm1,\,\pm2,\,\cdots \tag{4.50}$$

Any of these choices: 0 or $\pm k$, will provide the correct map in x_n so we select

$$\theta_{n+1} = 2\,\theta_n. \tag{4.51}$$

In terms of θ_0:

$$\theta_n = 2^n\,\theta_0. \tag{4.52}$$

We see that θ_n grows exponentially with n!

What does this mean for x_{n+1}? x_n is related to θ_n via the $\sin^2$, so the result is independent of the signs of θ_n. Moreover, computing a particular x_n is rooted in a choice of x_0 which corresponds to θ_0. Suppose we do this on a computer and further, assume θ_0 is represented as a binary number, then each iteration amounts to a simple "shift operation": a

multiplication by 2. So with each iteration, the right most digit in the computer is emptied. Inevitably, it is filled with some sort of numerical garbage. For example, suppose that a binary number under consideration is 1001 which can be written as

$$1001 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0. \tag{4.53}$$

Further suppose that the total digit of the computer is 4 (in any case, computer has finite digits and here we take 4 just for simplicity). Then an operation of multiplication by 2 on the above number results in

$$1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 = 1001g \tag{4.54}$$

where g represents garbage. The most left (highest) digit 1 is lost because the computer is assumed to have only 4 digits. Therefore, if a number is started with m digits it follows that after m iterations all information about the initial conditions is lost. The system has "forgotten" what it began from. This means that the value of θ_n is extremely sensitive to the initial value. This is one of the definitions of "Chaos in Dynamic Systems". In fact, say, we start from two initial conditions differed by $\epsilon > 0$ (as small as we wish)

$$\theta_0 - \bar{\theta}_0 = \epsilon\,, \tag{4.55}$$

then $\theta_n - \bar{\theta}_n = 2^n\,\theta_0 - 2^n\,\bar{\theta}_0 = 2^n\,(\theta_0 - \bar{\theta}_0) = 2^n\,\epsilon$ and so no matter how small ϵ is there will be an n large enough that this difference is of $O(1)$. For instance, for $\epsilon = 10^{-16}$, then it takes about 50 iterations for this to happen. Fig. 4.18 shows the chaotic orbit of the logistic map for the case $\mu = 4$ starting from the point $x_0 = 0.707$.

This shows a significant problem because all of the information about the sequence is contained in the initial conditions. The particular sequence you are on has lost its identity! It follows that the converse is true: if you start off with some initial condition, you cannot know where the sequence will take you, otherwise the sequence identity is preserved. This means the loss of the ability to predict, even though we have what is in principle a completely deterministic system, is lost: a profound result from a seemingly trivial quadratic operation!

We have seen from this example that a new possibility for the behavior of solutions of a discrete map arises. Formerly we had, for different initial conditions:

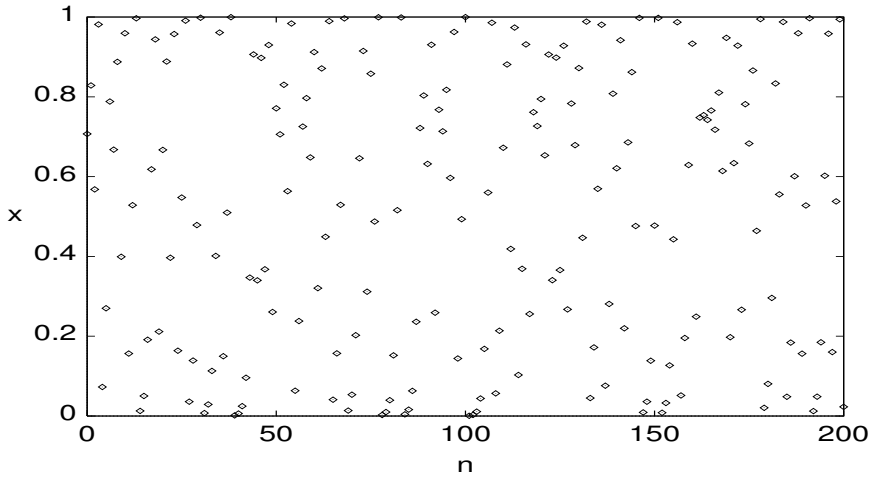


Fig. 4.18 The orbit of the logistic map when $\mu = 4$ and $x_0 = 0.707$.

(1) divergence to infinity;

(2) convergence to a single value;

(2) convergence to a cycle: oscillations.

Now the new 4th possibility is divergence but not to infinity: bounded divergent motion. This is **Chaos**. The ω-limit set for the case of $\mu = 4$ of the logistic map is all values of x_n.

Is this a paradox: "random determinism" or "deterministic noise" ?

Other properties of the logistic map are summarized in the logistic map picture, called the **orbit diagram**, a magnificent picture that has become an icon of nonlinear dynamics (Fig. 4.19). Fig. 4.19 plots the system's attractor as a function of μ. To generate the orbit diagram for yourself, you'll need to write a computer program with two "loops". First, choose a value of μ. Then generate an orbit starting from some random initial condition x_0. Iterate for 300 cycles or so, to allow the system to settle down to its eventual behavior. Once the transients have decayed, plot many points, say $x_{301}, \cdots, x_{600}$ above that μ. Then move to an adjacent value of μ and repeat, eventually sweeping across the whole picture.

Fig. 4.19 shows the most interesting part of the diagram, in the region $2.8 \leq \mu \leq 4$. At $\mu = 2.8$, the attractor is 1-cycle. As μ increases to $\mu = 3$, the 1-cycle bifurcates into a period-2 cycle, as indicated by the two branches. As μ further increases, both branches split simultaneously, yielding a period-4 cycle. This splitting is the period-doubling bifurcation

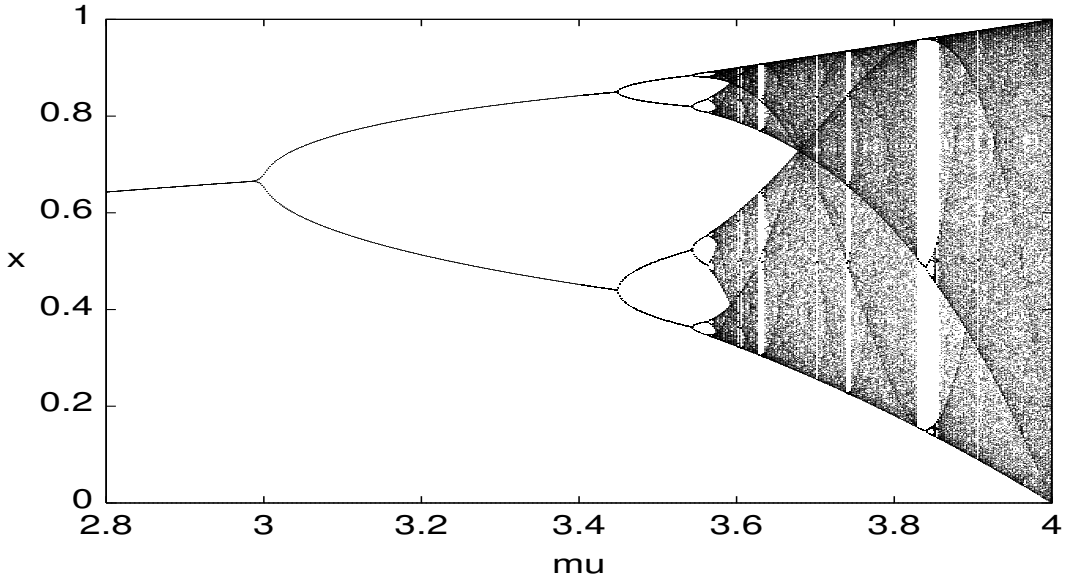


Fig. 4.19 The orbit diagram of the logistic map.

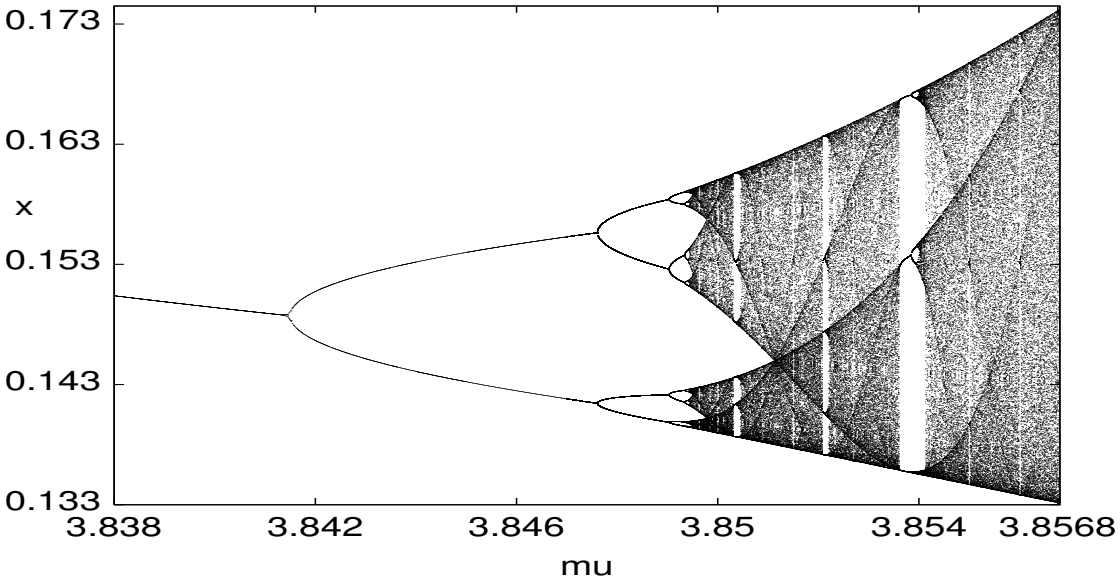


Fig. 4.20 A blow up of part of the period-3 window of the logistic map.

mentioned earlier. A cascade of further period-doubling occurs as μ increases, yielding period-8, period-16, and so on, until at $\mu = \mu_\infty \approx 3.570$, the map becomes chaotic and the attractor changes from a finite to an infinite set of points.

For $\mu > \mu_\infty$ the orbit diagram reveals an unexpected mixture of order and chaos, with *periodic windows* interspersed between chaotic clouds of dots. The large window beginning

near $\mu \approx 3.85$ contains a stable period-3 cycle. A blow-up of part of the period-3 window (Fig. 4.20) is shown in the lower panel of Fig. 4.19. Fantastically, a copy of the orbit diagram reappears in miniature!

4.7 Divergence of Adjacent Solutions: Liapunov Exponents

On the logistic map the region of "chaos" actually consists of a series of attracting chaotic and regular motion (intermittency) (see Fig. 4.19). Now we introduce new concepts which are more suitable to define chaotic dynamics.

We have seen that the logistic map can exhibit aperiodic orbits for certain parameter values, but how do we know that this is really chaos? To be called "chaotic," a system should also show *sensitive dependence on initial conditions*, in the sense that neighboring orbits separate exponentially fast, on average. Here is the intuition of the Liapunov exponents. Given some initial condition x_0, consider a nearby point $x_0+\delta_0$, where the initial separation δ_0 is extremely small. Let δ_N be the separation after N iterates. If $|\,\delta_N\,| \approx |\,\delta_0\,|\, e^{N\lambda}$, then λ is called the Liapunov exponent. A positive Liapunov exponent is a signature of chaos.

A more precise and computationally useful formula for λ can be derived. By taking logarithms and noting that $\delta_N = f^N(x_0+\delta_0) - f^N(x_0)$, we obtain

$$\begin{aligned}
\lambda(x_0) &\approx \frac{1}{N}\ln\left|\frac{\delta_N}{\delta_0}\right| \\
&= \frac{1}{N}\ln\left|\frac{f^N(x_0+\delta_0)-f^N(x_0)}{\delta_0}\right| \\
&= \frac{1}{N}\ln\left|(f^N)'(x_0)\right| \qquad \text{as } \delta_0 \to 0,
\end{aligned} \tag{4.56}$$

where we've taken the limit $\delta_0 \to 0$ in the last step. The term inside the logarithm can be expanded by the chain rule:

$$(f^N)'(x_0) = \prod_{i=0}^{N-1} f'(x_i)\,.$$

Hence,

$$\lambda(x_0) \approx \frac{1}{N}\ln\left|\prod_{i=0}^{N-1} f'(x_i)\right| = \frac{1}{N}\sum_{i=0}^{N-1}\ln|\,f'(x_i)\,|\,. \tag{4.57}$$

Note that $\lambda(x_0)$ depends on x_0. So what value of n is universal "irrespective" of where x_0 happens to be? For a non empty ω-limit set, the answer is $n \to \infty$. This is because for

any initial condition, after enough long (finite) transient period, there are infinite iterations on the ω-limit set. Thus,,

$$\lambda = \lim_{n\to\infty} \left\{ \frac{1}{N} \sum_{i=0}^{N-1} \ln | f'(x_i) | \right\}. \tag{4.58}$$

which is the same for all x_0 in the basin of attraction of a given attractor. The λ is called **liapunov exponent** or **Liapunov coefficient**. For stable fixed points and cycles, λ is negative; for chaotic attractors, λ is positive. The positive λ characterizes *local* divergence. It does *not* say whether or not $x_n \to \infty$ or not alternatively if $\lambda > 0$. If $\lambda > 0$ and the motion is bounded in "one" region, then we have chaos. Some authors would also requires that ω-limit set be "dense". That is the case here for $\mu = 4$. ω-limit set is a "line" instead of a bunch of points. Note that this implies that the iterates "fill up" space: e.g. for $\mu = 3.9$. See the cobweb diagram in Fig. 4.21.

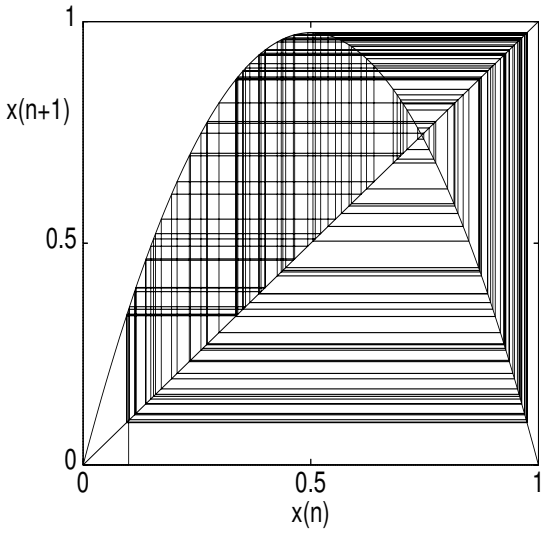


Fig. 4.21 Cobweb diagram when $\mu = 3.9$.

The next two examples deal with special cases where λ can be found analytically.

EXAMPLE 4.1. Suppose that f has a stable p-cycle containing the point x_0. Show that the Liapunov exponent $\lambda < 0$. If the cycle is superstable, show that $\lambda = -\infty$.

Solution: As usual, we convert questions about p-cycles of f into questions about fixed points of f^p. Since x_0 is an element of a p-cycle, x_0 is a fixed point of f^p. By assumption,

the cycle is stable; hence the multiplier $|\,(f^p)'(x_0)\,| < 1$. Therefore, $\ln|\,(f^p)'(x_0)\,| < \ln(1) = 0$, a result that we'll use in a moment.

Next, observe that for a p-cycle,

$$\lambda = \lim_{n\to\infty}\left\{\frac{1}{N}\sum_{i=0}^{N-1}\ln|\,f'(x_i)\,|\right\} = \frac{1}{p}\sum_{i=0}^{N-1}\ln|\,f'(x_i)\,|$$

since the same p terms keep appearing in the infinite sum. Finally, using the chain rule in reverse, we obtain

$$\lambda = \frac{1}{p}\sum_{i=0}^{N-1}\ln|\,f'(x_i)\,| = \frac{1}{p}\ln|\,(f^p)'(x_0)\,| < 0\,,$$

as desired. If the cycle is superstable, then $|\,(f^p)'(x_0)\,| = 0$ by definition, and then $\lambda = \frac{1}{p}\ln(0) = -\infty$.

EXAMPLE 4.2. Find the Liapunov exponent of the **tent map**, defined by

$$f(x) = \begin{cases} \mu\,x\,, & 0 \le x \le \frac{1}{2} \\ \mu - \mu\,x\,, & \frac{1}{2} \le x \le 1 \end{cases} \tag{4.59}$$

for $0 \le \mu \le 2$ and $0 \le x \le 1$ (Fig. 4.22).

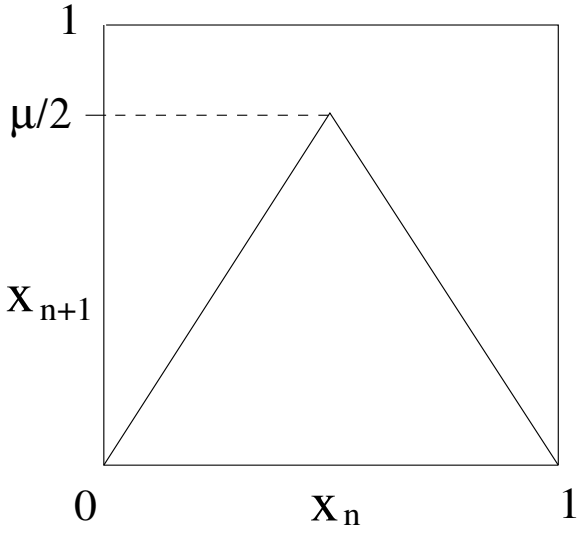


Fig. 4.22 The tent map.

Solution: Because it is piecewise linear, the tent map is far easier to analyze than the logistic map. Since $f'(x) = \pm\,\mu$ for $[0,\,\frac{1}{2}) \cup (\frac{1}{2},\,1]$, thus, $\lambda = \lim_{N\to\infty}\left\{\frac{1}{N}\sum_{i=0}^{N-1}\ln|\,f'(x_i)\,|\right\} = \ln\mu\,.$

Example 4.2 suggests that the tent map has chaotic solutions for all $\mu > 1$, since $\lambda = \ln \mu > 0$. In fact, the dynamics of the tent map can be understood in detail, even in the chaotic regime; see Devaney (1989).

In general, one needs to use a computer to calculate Liapunov exponents. The next example outlines such a calculation for the logistic map.

EXAMPLE 4.3. Describe a numerical scheme to compute λ for the logistic map

$$x_{n+1} = \mu\, x_n\, (1 - x_n).$$

Graph the results as a function of the control parameter μ, for $2.8 \leq \mu \leq 4$.

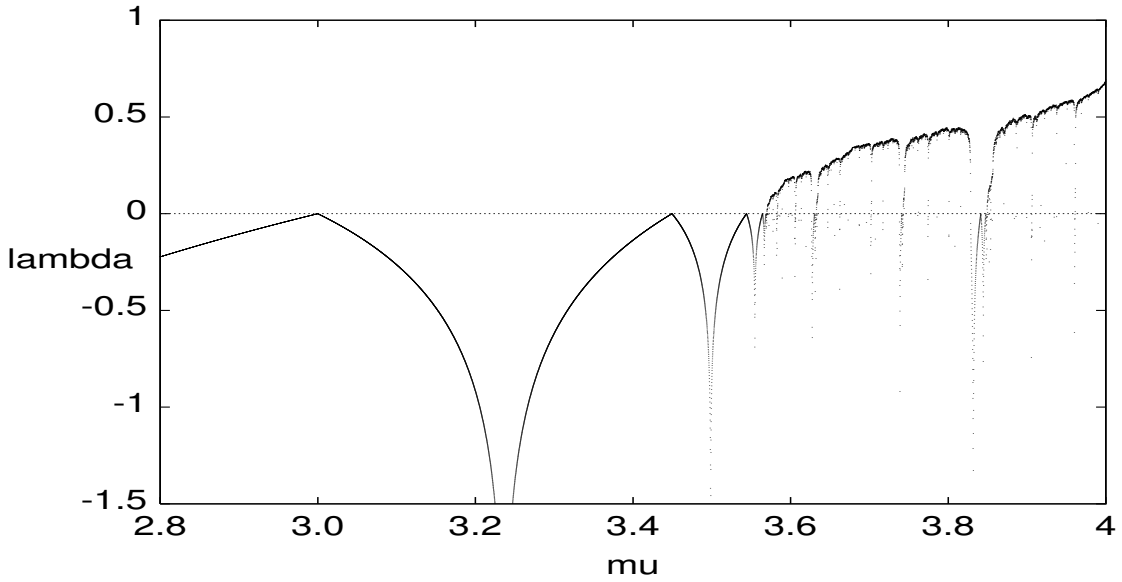


Fig. 4.23 The Liapunov exponents of the logistic map.

Solution: Fix some value of μ. Then starting from a random initial condition, iterate the map long enough to allow transients to decay, say, 300 iterates or so. Next compute a large number of additional iterates, say, 10,000. You only need to store the current value of x_n, not all the previous iterates. Compute $\ln |\, f'(x_n)\,| = \ln |\, \mu\,(1 - 2\,x_n)\,|$ and add it to the sum of the previous logarithms. The Liapunov exponent is then obtained by dividing the grand total by 10,000. Repeat this procedure for the next μ, and so on. The end result should look like Fig. 4.23.

4.8 Intermittency: Route to Chaos

There are other routes to chaos. One can have a particular μ which seems to produce nonchaotic behavior, then abruptly bursts of chaos will set and complete chaos will set in similarly.

Transients are the issue here. Another slightly different phenomenon involves *spontaneous transitions between seemingly clam states*, like gusts of winds but not to be a different state afterwards. These transitions are like events without external causes. In the logistic map they can occur near $\mu \approx 1 + 2\sqrt{2}$. (See Fig. 4.24.) This corresponds to a "window" in the ω-limit set chaotic region where the Liapunov coefficient becomes briefly negative. The chaotic system does a rapid period doubling identical in character to what happened previously as μ increases. As μ decreases from the 3-cycle, the reverse happens. The periodic behavior lasts for shorter intervals until it disappears entirely. This is another entirely different approach to fill chaos than the period-doubling. (There are others not listed here.)

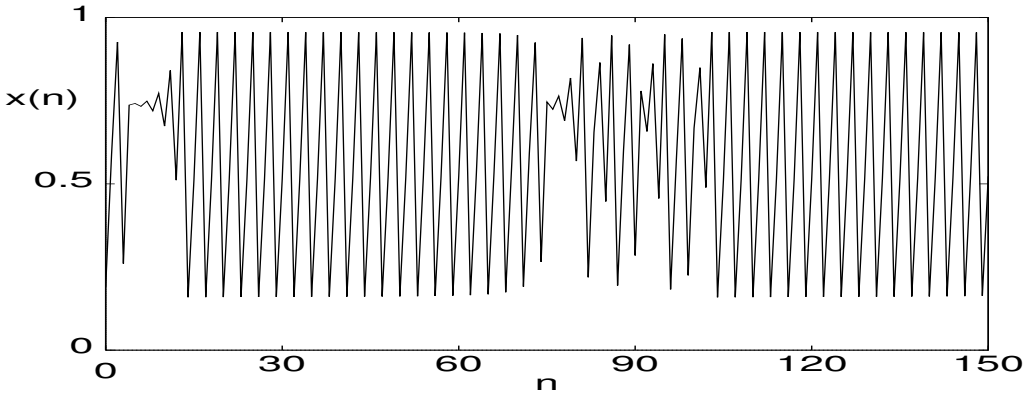


Fig. 4.24 Intermittency to chaos.

Chapter 5

BIFURCATIONS IN CONTINUOUS DIFFERENTIAL EQUATIONS

We have seen bifurcations in a discrete map. There are many types of bifurcations associated with differential equations and maps. It is through changes in qualitative behavior of this type that complex behavior such as chaos arise. It is thus worthwhile to understand the qualitative nature of these phenomena in differential equations.

In this chapter we describe the simplest but most fundamental bifurcations of equilibria. These are represented by the following three differential equations which depend on a single parameter μ:

$$\dot{x} = \mu - x^2 \qquad (saddle - node) \tag{5.1}$$

$$\dot{x} = \mu\, x - x^2 \qquad (transcritical) \tag{5.2}$$

$$\dot{x} = \mu\, x - x^3 \qquad (pitchfork). \tag{5.3}$$

5.1 Saddle-Node Bifurcation

Consider the nonlinear ODE (5.1)

$$\dot{x} = \mu - x^2$$

where μ is parameter (also called control parameter). We note that it will have fixed points when $\dot{x} = 0$, or $x^2 = \mu$. In particular, as μ changes from negative to positive we first have no fixed points, then one and then two fixed points. This is depicted in Fig. 5.1.

We note as with the logistic equation that the fixed points have stable and unstable properties depending upon the circumstances. For the DE (5.1), we see that for $\mu > 0$, there are two fixed points $x^* = x_\pm = \pm\sqrt{\mu}$. To determine the stability of x_+, let

$$x = x_+ + \eta$$

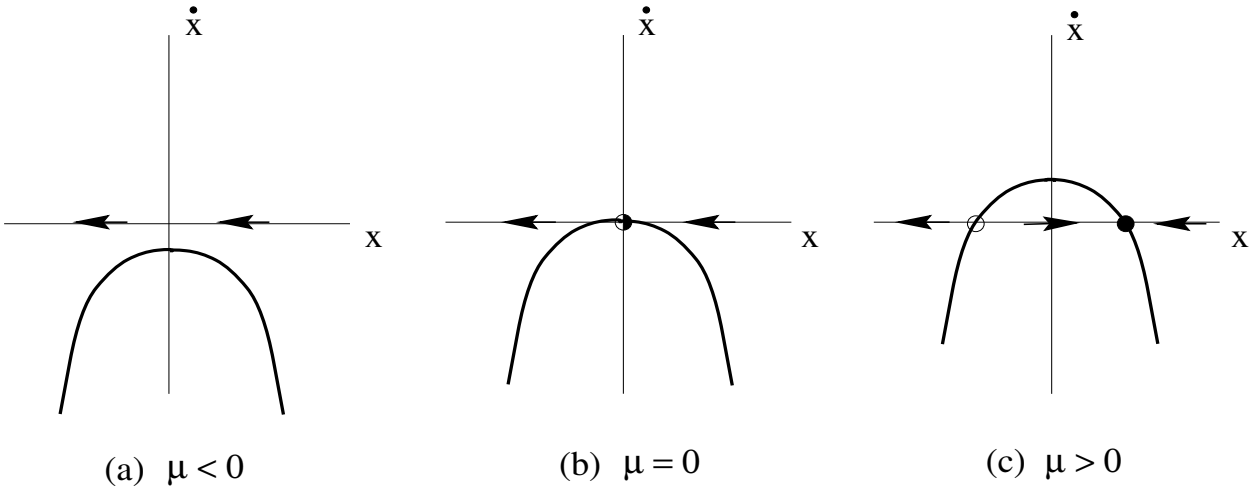


Fig. 5.1 Fixed points of saddle-node bifurcation and their stability.

where η is a small perturbation (disturbance). Thus,

$$\begin{aligned} \dot{x} = \dot{x}_+ + \dot{\eta} &= \mu - x^2 \\ &= \mu - (x_+ + \eta)^2 \\ &= \mu - x_+^2 - 2\,x_+\,\eta - \eta^2. \end{aligned} \tag{5.4}$$

Since x_+ satisfies equation (5.1), so $\dot{x}_+ = \mu - x_+^2 = 0$, hence, (5.4) becomes

$$\dot{\eta} = -\,2\,x_+\,\eta - \eta^2 = -\,2\,x_+\,\eta + O(\eta^2) \approx -\,2\,x_+\,\eta. \tag{5.5}$$

The solution η is given by

$$\eta = C\,e^{-\,2\,x_+\,t} \qquad (C \neq 0). \tag{5.6}$$

Since $x_+ > 0$, thus $|\eta| \to 0$ as $t \to \infty$. So the nearby points of x_+ move towards this fixed point x_+. Hence, x_+ is stable.

Note from (5.5) that since $x_+ > 0$, so for small η, $\eta > 0 \Longrightarrow \dot{\eta} < 0 \Longrightarrow \dot{x} = \dot{x}_+ + \dot{\eta} = \dot{\eta} < 0$ ($\dot{x}_+ = 0$). Similarly, $\eta < 0 \Longrightarrow \dot{x} > 0$. That is, $\dot{x}$ is positive to the left and negative to the right.

For the other fixed point $x_- = -\sqrt{\mu}$, the sign is reverse, so x_- is unstable when $\mu > 0$.

From the above perturbation analysis, we have a simple way to determine the stability

of fixed points. Suppose x^* is a fixed point of the DE $\dot{x} = f(x)$, i.e., $f(x^*) = 0$. Then

$$\text{Fixed point } x^* \begin{cases} \text{is stable} & \text{if } f'(x^*) < 0; \\ \text{is unstable} & \text{if } f'(x^*) > 0. \end{cases} \tag{5.7}$$

For equation (5.1), since $f'(x^*) = -2x^*$, so $f'(x_+) = -2x_+ < 0$ and $f'(x_-) = -2x_- > 0$ thus, x_+ is stable and x_- is unstable. The above discussion is illustrated in Fig. 5.1c.

For $\mu = 0$, we have $\dot{x} = -x^2$. We see that $\dot{x} < 0$ for $x \neq 0$, so it is stable for $x > 0$ and is unstable for $x < 0$. It is called *half-stable* fixed point (Fig. 5.1b). The point defined by $\mu_c = 0$ is called *critical point* or *bifurcation point*, at which the function $f(x)$ has a repeated roots $x = 0$. In general, the bifurcation points are determined from the equations $f(x) = 0$ and $f'(x) = 0$. For example, consider (5.1), $f'(x) = -2x = 0$ gives us $x_c = 0$ and then from $f(x_c) = \mu - x_c^2 = \mu = 0$ we obtain the critical point $\mu_c = 0$.

The relation between the parameter μ and the fixed points x^* is shown in Fig. 5.2. Note that the two fixed points coalesce at the bifurcation point $\mu = 0$ at which $f(x^*) = 0$. This figure is called bifurcation diagram.

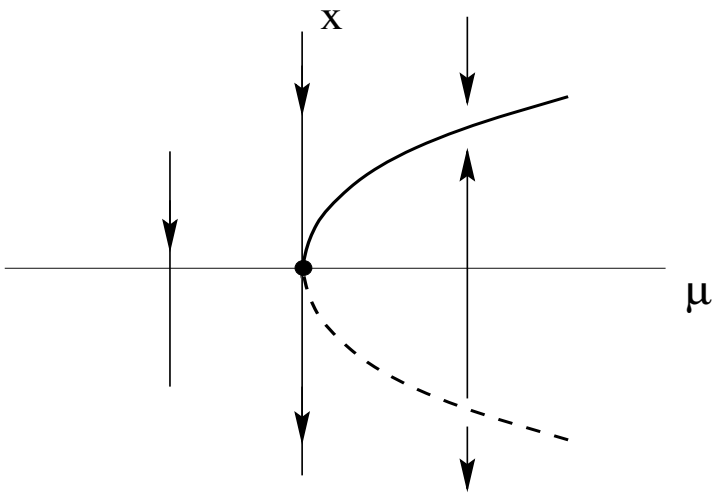


Fig. 5.2 Bifurcation diagram for saddle-node bifurcation, —— stable; - - - - unstable.

The saddle-node bifurcation is sometimes called a *fold bifurcation* (because the curve in Fig. 5.2 has a fold in it) or a *turning-point bifurcation* (because the point $(x, \mu) = (0, 0)$ is a "turning point"). Admittedly, the term *saddle-node* doesn't make much sense for one dimensional vector field (differential equation). The name derives from higher-dimensional

context, such as vector fields on the plane as we will seen in Chapter 7.

EXAMPLE 5.1. Find the fixed points of the DE

$$\dot{x} = 1 + \mu\, x + x^2 \tag{5.8}$$

and study their stability.

Solution: Here $f(x) = 1 + \mu\, x + x^2$. The fixed points are determined from $f(x) = 0$; i.e.,

$$1 + \mu\, x + x^2 = 0 \qquad \Longrightarrow \qquad x^* = x_\pm = \frac{1}{2}\left[-\mu \pm \sqrt{\mu^2 - 4} \right] \tag{5.9}$$

Thus, when $\mu < -2$ or $\mu > 2$, there are two fixed points; when $\mu = -2$ or $\mu = 2$, there is one fixed point; and when $-2 < \mu < 2$ there is no fixed point. The stability is determined from the derivative $f'(x) = \mu + 2\, x$. So

$$f'(x_\pm) = \mu + 2 \times \frac{1}{2}\left[-\mu \pm \sqrt{\mu^2 - 4} \right] = \pm \sqrt{\mu^2 - 4} \tag{5.10}$$

which indicates that $f'(x_+) > 0$ and $f'(x_-) < 0$. Hence, x_+ is unstable and x_- is stable.

When $\mu = \pm 2$, $x = -\dfrac{\mu}{2} = \mp 1$. This is a bifurcation point, and stability is not determined at this point. It should be noted that there are two separate saddle-node bifurcations, shown in Fig. 5.3.

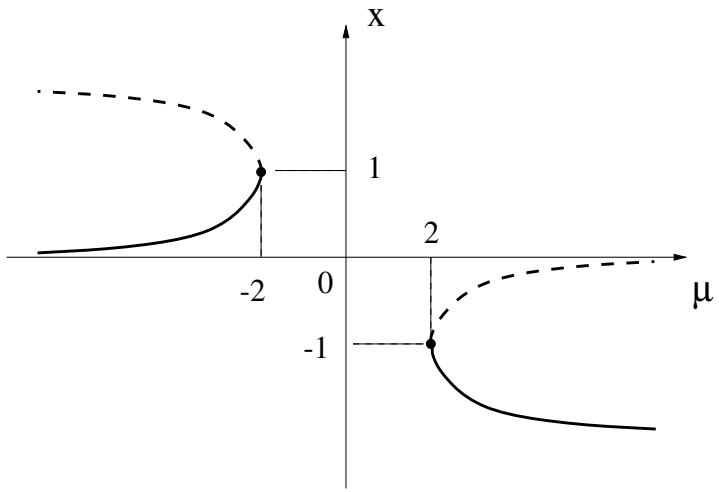


Fig. 5.3 Bifurcation diagram for Example 5.1, —— stable; - - - - unstable.

From the viewpoint of stability analysis, sometimes it is convenient to use a graph approach since explicit solutions of fixed points are not easy to be found for a general

nonlinear equation. Let's consider the equation (5.1) again. Think of the equation as a family of D.E. and so the vector field is one dimension (on a line). Use the information from $\dot{x}$ to find whether x increases (moving to right) or decreases (moving to left). As we discussed above, when μ is positive, there are two fixed points, one stable and one unstable (Fig. 5.1c).

As μ approaches 0 from above, the parabola moves down and the two fixed points move toward each other. When $\mu = 0$, the fixed points coalesce into a "half-stable" fixed point at $x = 0$ (Fig. 5.1b). This type of fixed point is extremely delicate – it vanishes as soon as $\mu < 0$, and now there are no fixed points at all (Fig. 5.1a). Note that the arrows shown in Fig. 5.1 indicate the moving direction of x. They are determined by the sign of $\dot{x}$. If $\dot{x} < 0$ (> 0), the direction is from right to left (from left to right). Let's use the graph approach to consider the following example.

EXAMPLE 5.2. Show that the first-order system

$$\dot{x} = \mu - x - e^{-x} \tag{5.11}$$

undergoes a saddle-node bifurcation as μ is varied, and find the value of μ at the bifurcation point.

Solution: The fixed points satisfy $f(x, \mu) = \mu - x - e^{-x} = 0$. But now we run into a difficulty - we can't find the fixed points explicitly as a function of μ. Instead we adopt a geometric approach. One method is to graph the function f for different values of μ, look for its roots x^*, and then sketch the vector field on the x-axis. This method is fine, but there is an easier way. The point is that the two functions $\mu - x$ and e^{-x} have much more familiar graphs than their difference $\mu - x - e^{-x}$. So we plot $\mu - x$ and e^{-x} on the same picture (Fig. 5.4), where the line $\mu - x$ intersects the curve e^{-x}, we have $\mu - x = e^{-x}$ and so $f(x, \mu) = 0$. *Thus, intersections of the line and the curve correspond to fixed points for the system.* This picture also allows us to read off the direction of flow on the x-axis: the flow is to the right where the line lies above the curve, since $\mu - x > e^{-x}$ and therefore $\dot{x} > 0$. Hence, the fixed point on the right is stable, and the one on the left is unstable.

Now imagine we start decreasing the parameter μ. The line $\mu - x$ slides down and the

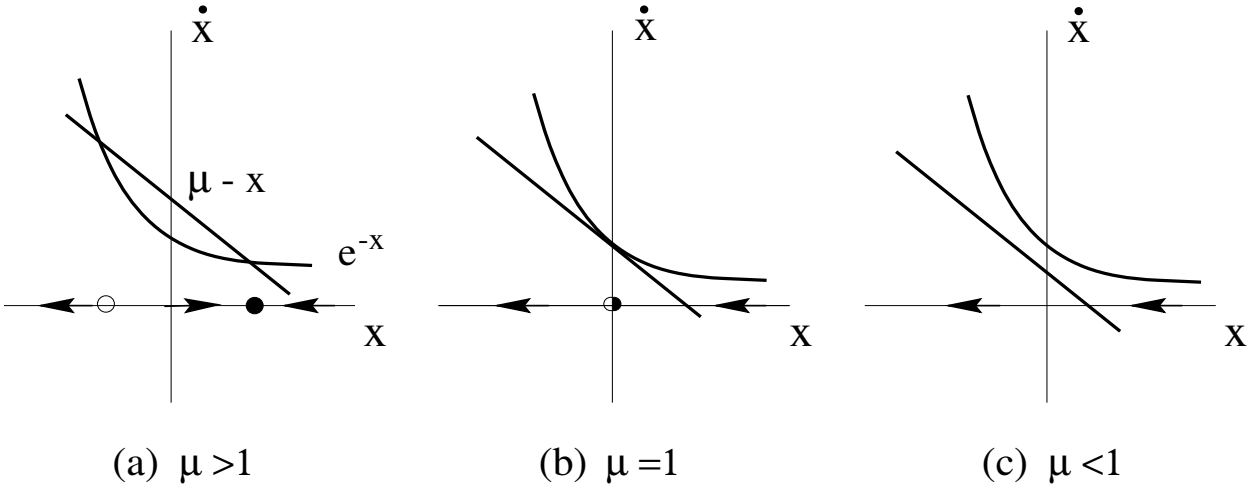


Fig. 5.4 Fixed points and bifurcation for Example 5.2.

fixed points approach each other. At some critical value $\mu = \mu_c$, the line becomes *tangent* to the curve and the fixed points coalesce in a saddle-node bifurcation (Fig. 5.4b). For μ below this critical value, the line lies below the curve and there are no fixed points (Fig. 5.4c).

To find the bifurcation point $\mu = \mu_c$ we impose the condition that the graphs of $\mu - x$ and e^{-x} intersect *tangentially*. Thus we demand equality of the functions *and* their derivatives:

$$e^{-x} = \mu - x \qquad \text{and} \qquad \frac{d}{dx}e^{-x} = \frac{d}{dx}(\mu - x)\,. \tag{5.12}$$

The second equation implies $-e^{-x} = -1$, so $x = 0$. Then the first equation yields $\mu = 1$. Hence, the bifurcation point is $\mu_c = 1$, and the bifurcation occurs at $x = 0$.

Let's use both graphical and analytical approaches to consider one more example.

EXAMPLE 5.3. Consider the DE

$$\dot{x} = \mu + \frac{1}{2}\,x - \frac{x}{1+x}. \tag{5.13}$$

Find the critical value of μ and sketch the bifurcation diagram in the $(x,\, \mu)$ plane.

Solution: Equation (5.13) can be rewritten as

$$\dot{x} = (\mu - 1) + \frac{1}{2}\,x - \frac{-1}{1+x} = f(x) - g(x) \quad \text{with } f(x) = (\mu - 1) + \frac{1}{2}\,x \text{ and } g(x) = \frac{-1}{1+x} \tag{5.14}$$

where $f(x)$ is a straight line while $g(x)$ represents a simple hyperbola. Thus we can use graphic approach to find possible bifurcations and qualitatively different vector fields as μ varied. Note that we need to consider two subintervals $x \in (-\infty, -1)$ and $x \in (-1, \infty)$ separately since the vector field is discontinuous at $x = -1$. The bifurcation points are determined by

$$f(x) = g(x) \qquad \text{and} \qquad f'(x) = g'(x) \tag{5.15}$$

or

$$(\mu - 1) + \frac{1}{2}\, x = \frac{-1}{1+x} \qquad \text{and} \qquad \frac{1}{2} = \frac{1}{(1+x)^2} \tag{5.16}$$

which gives two critical (bifurcation) points:

$$\begin{aligned} x &= -1 - \sqrt{2}\,, \quad \mu = \frac{3}{2} + \sqrt{2} \qquad \text{for } x < -1\,; \\ x &= -1 + \sqrt{2}\,, \quad \mu = \frac{3}{2} - \sqrt{2} \qquad \text{for } x > -1. \end{aligned} \tag{5.17}$$

Thus, the qualitatively behavior of the DE (5.13) for $x < -1$ can be obtained as illustrated in Fig. 5.5. It is a saddle-node bifurcation. Similarly, one may obtain similar diagrams for $x > -1$. The stability of the fixed points can be easily found from Figs. 5.5 and 5.6: for $x < -1$, x_- is unstable while x_+ is stable; for $x > -1$, x_- is stable while x_+ is unstable.

For this question, however, we can find the fixed points and their stability conditions analytically though some algebra is needed. First, the fixed points can be solved from $\dot{x} = 0$, i.e.,

$$x^2 + (2\,\mu - 1)\, x + 2\,\mu = 0 \tag{5.18}$$

which yields fixed points:

$$x^* = x_\pm = \frac{1}{2}\left[1 - 2\,\mu \pm \sqrt{(2\,\mu - 3)^2 - 8}\right]. \tag{5.19}$$

It is seen from equation (5.13) that $x \neq -1$, i.e., either $x < -1$ or $x > -1$. To have real solutions for the fixed points x^*, the following condition

$$(2\,\mu - 3)^2 - 8 \geq 0 \qquad \text{or} \qquad \left(\mu - \frac{3}{2}\right)^2 \geq 2 \tag{5.20}$$

must be satisfied, i.e.

$$\mu \geq \frac{3}{2} + \sqrt{2} \qquad \text{or} \qquad \mu \leq \frac{3}{2} - \sqrt{2}\,. \tag{5.21}$$

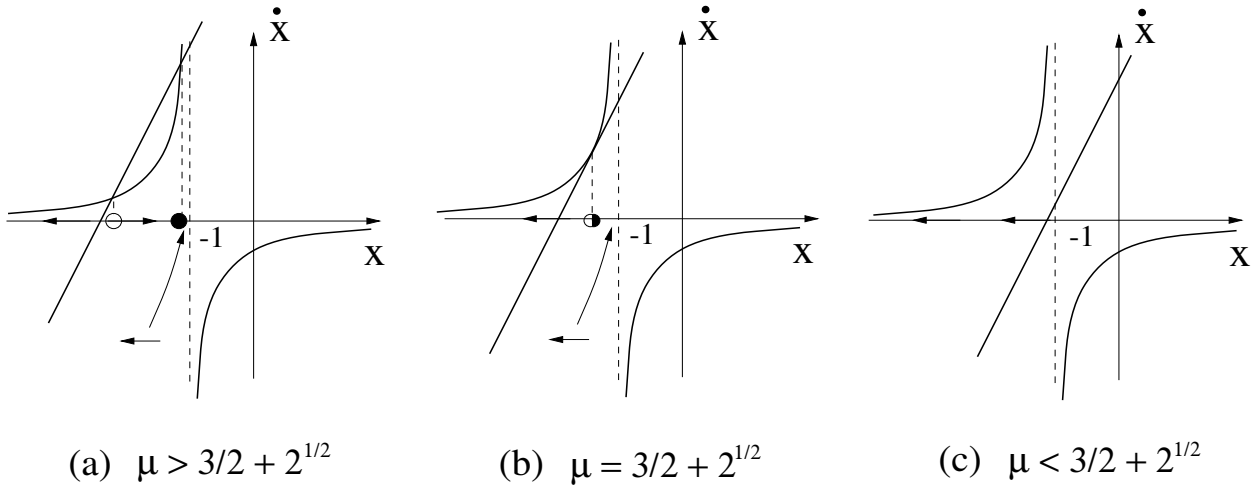


Fig. 5.5 Fixed points of system (5.13) when $x < -1$.

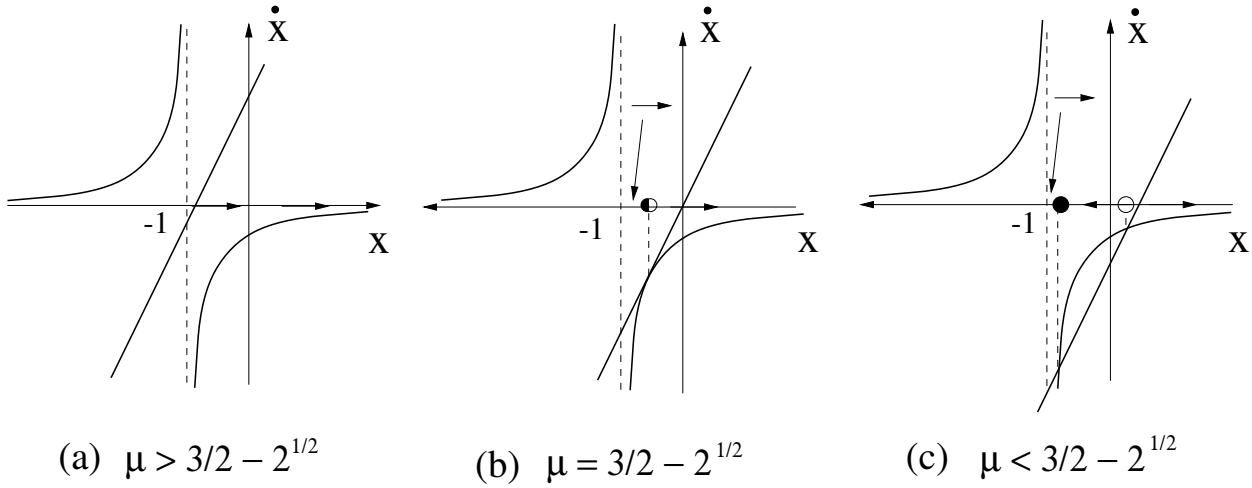


Fig. 5.6 Fixed points of system (5.13) when $x > -1$.

From the equilibrium equation $\dot{x} = 0$, we have

$$\mu = \frac{1}{2} - \frac{1}{2}\,x - \frac{1}{1+x}\,. \tag{5.22}$$

So, when $x < -1$, $\mu > \dfrac{1}{2} - \dfrac{1}{2}\,x > \dfrac{1}{2} - \dfrac{1}{2}\,(-1) = 1$, which indicates that the fixed points $x_{\pm}$ corresponding to $\mu \geq \dfrac{3}{2} + \sqrt{2}$ must belong to the interval $x < -1$. Similarly, the fixed

points $x_\pm$ corresponding to $\mu \le \frac{3}{2} - \sqrt{2}$ must located in the interval $x > -1$. Therefore,

$$\begin{aligned}
&\text{when} \quad \mu > \frac{3}{2} + \sqrt{2} \quad \Longrightarrow \quad \text{two fixed points } x_\pm < -1 \text{ (given by (5.19))} \\
&\text{when} \quad \mu = \frac{3}{2} + \sqrt{2} \quad \Longrightarrow \quad \text{one (repeated) fixed point } x^* = -1 - \sqrt{2} \\
&\text{when} \quad \mu < \frac{3}{2} - \sqrt{2} \quad \Longrightarrow \quad \text{two fixed points } x_\pm > -1 \text{ (given by (5.19))} \\
&\text{when} \quad \mu = \frac{3}{2} - \sqrt{2} \quad \Longrightarrow \quad \text{one (repeated) fixed point } x^* = -1 + \sqrt{2} \\
&\text{when} \quad \frac{3}{2} - \sqrt{2} < \mu < \frac{3}{2} + \sqrt{2} \quad \Longrightarrow \quad \text{no fixed points.}
\end{aligned} \tag{5.23}$$

The stability of $x_\pm$ can be determined from

$$f'(x_\pm) = \frac{1}{2} - \frac{1}{(1+x_\pm)^2}\,. \tag{5.24}$$

Since

$$\begin{aligned}
(1+x_\pm)^2 &= \left\{1 + \frac{1}{2}\left[1 - 2\,\mu \pm \sqrt{(2\,\mu-3)^2 - 8}\right]\right\}^2 \\
&= \left\{\frac{1}{2}\left[3 - 2\,\mu \pm \sqrt{(2\,\mu-3)^2 - 8}\right]\right\}^2 \\
&= \left[\mu - \frac{3}{2} \mp \sqrt{(\mu - \frac{3}{2})^2 - 2}\right]^2
\end{aligned} \tag{5.25}$$

then

$$\begin{aligned}
\frac{d}{d\mu}\left[(1+x_\pm)^2\right] &= 2\left[\mu - \frac{3}{2} \mp \sqrt{(\mu - \frac{3}{2})^2 - 2}\right]\left[1 \mp \frac{\mu - \frac{3}{2}}{\sqrt{(\mu - \frac{3}{2})^2 - 2}}\right] \\
&= \frac{\mp 2\left[\mu - \frac{3}{2} \mp \sqrt{(\mu - \frac{3}{2})^2 - 2}\right]^2}{\sqrt{(\mu - \frac{3}{2})^2 - 2}}\,.
\end{aligned} \tag{5.26}$$

Therefore, for the fixed point x_+, $\frac{d}{d\mu}\left[(1+x_+)^2\right] < 0$ for $\mu \in (-\infty, \infty)$, i.e., $(1+x_+)^2$ is monotonically decreasing when μ increases, so

$$\begin{aligned}
&\text{if} \quad \mu < \frac{3}{2} - \sqrt{2} \quad \Longrightarrow \quad (1+x_+)^2 > \left[1 - (-1+\sqrt{2})\right]^2 > 2 \quad \Longrightarrow \quad f'(x_+) > 0\,; \\
&\text{if} \quad \mu > \frac{3}{2} + \sqrt{2} \quad \Longrightarrow \quad (1+x_+)^2 < \left[1 - (-1-\sqrt{2})\right]^2 < 2 \quad \Longrightarrow \quad f'(x_+) < 0\,;
\end{aligned} \tag{5.27}$$

which indicates that

$$\begin{aligned} &\text{when} \quad \mu < \frac{3}{2} - \sqrt{2}\,, \quad x_+ \ \text{is unstable;} \\ &\text{when} \quad \mu > \frac{3}{2} + \sqrt{2}\,, \quad x_+ \ \text{is stable.} \end{aligned} \tag{5.28}$$

Similarly, for the fixed point x_-, $\dfrac{d}{d\mu}\left[(1+x_-)^2\right] > 0$ for $\mu \in (-\infty, \infty)$, i.e., $(1+x_-)^2$ is monotonically increasing when μ increases, so we have

$$\begin{aligned} &\text{if} \quad \mu < \frac{3}{2} - \sqrt{2} \Longrightarrow (1+x_-)^2 < \left[1-(-1+\sqrt{2})\right]^2 < 2 \Longrightarrow f'(x_-) < 0 \Longrightarrow x_- \ \text{is stable;} \\ &\text{if} \quad \mu > \frac{3}{2} + \sqrt{2} \Longrightarrow (1+x_-)^2 > \left[1-(-1-\sqrt{2})\right]^2 > 2 \Longrightarrow f'(x_-) > 0 \Longrightarrow x_- \ \text{is unstable.} \end{aligned} \tag{5.29}$$

The bifurcation diagram can be determined from the equation $\mu = 1 - \dfrac{1}{2}x - \dfrac{1}{1+x}$ (Fig. 5.7). It has two branches, both are saddle-node bifurcations, as expected.

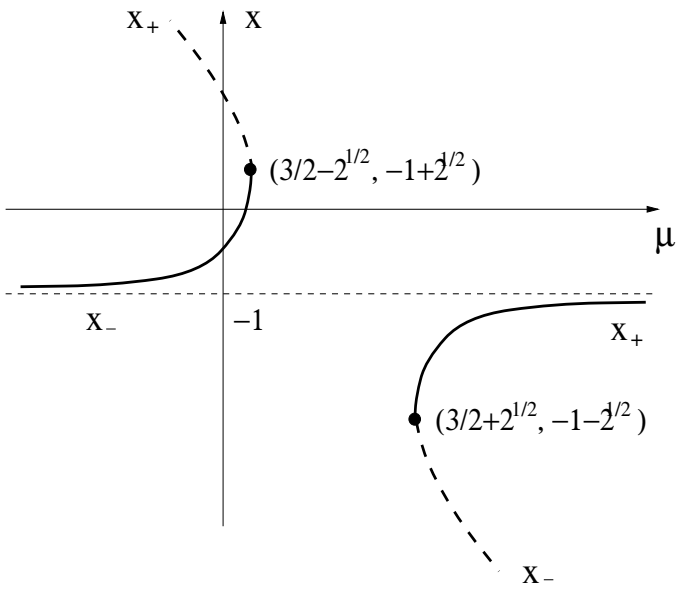


Fig. 5.7 Bifurcation diagram of system (5.13).

The approximation of the bifurcation diagram near the critical points can be found from the Taylor series:

$$\begin{aligned} \mu &= 1 - \frac{1}{2}x - \frac{1}{1+x} = \frac{5}{2} - \frac{7}{4}\sqrt{2} + \left(1 - \frac{\sqrt{2}}{2}\right)x - \frac{\sqrt{2}}{4}x^2 + O(x^3) \\ &\approx \left(\frac{3}{2} - \sqrt{2}\right) - \frac{\sqrt{2}}{4}\left(x + 1 - \sqrt{2}\right)^2 \qquad \text{for} \quad x < -1 \end{aligned}$$

$$\begin{aligned} \mu &= 1 - \frac{1}{2}\,x - \frac{1}{1+x} = \frac{5}{2} + \frac{7}{4}\sqrt{2} + \left(1 + \frac{\sqrt{2}}{2}\right) x + \frac{\sqrt{2}}{4}\,x^2 + O(x^3) \\ &\approx \left(\frac{3}{2} + \sqrt{2}\right) + \frac{\sqrt{2}}{4}\left(x + 1 + \sqrt{2}\right)^2 \qquad \text{for } \; x > -1, . \end{aligned} \tag{5.30}$$

The differential equation (5.1)

$$\dot{x} = \mu - x^2$$

is known as a **normal form** for this type of bifurcation as even for a general DE $\dot{x} = f(x)$. See Fig. 5.8, where the function $f(x)$ looks a parabola when it is close enough to the line $\dot{x} = 0$.

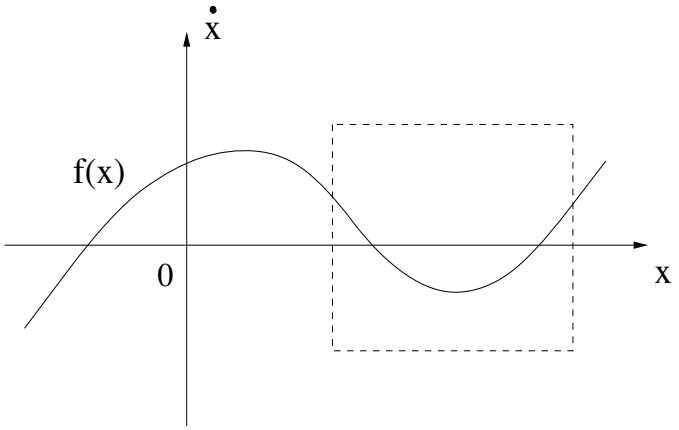


Fig. 5.8 Normal form of the saddle-node bifurcation.

The behavior of $f(x)$ near the line $\dot{x} = 0$ acts the same way as the normal form when a parameter is varied. We can show by a Taylor expansion that we are treating the same scenario. Normal form theory is very useful for the analysis of local behavior of nonlinear dynamics. The basic idea is to use a sequence of nonlinear coordinate transformations to transform the DE $\dot{x} = f(x)$ (which may be in vector form) to a new system

$$\dot{y} = g(y) \tag{5.31}$$

"simple as much as possible". Suppose $f(x)$ is given in Taylor series, then "simple as much as possible" means that *all inessential terms* in the Taylor series have been removed. But the resulting DE (5.31) keeps *qualitative* (or *topological*) behavior unchanged.

5.2 Transcritical Bifurcation

The importance of the saddle-node bifurcation is that all bifurcations of one-parameter families at an equilibrium with a zero eigenvalue can be perturbed to saddle-node bifurcations. Thus one expects that the zero eigenvalue bifurcations encountered in applications will be saddle-nodes. If they are not, then there is probably something special about the formulation of the problem which restricts the context so as to prevent the saddle-node from occurring. The *transcritical* bifurcation is one example which illustrates how the setting of the problem can rule out the saddle-node bifurcation.

There are certain scientific situations where an equilibrium (fixed point) must exist for all values of a parameter and can never be destroyed. For example, in the logistic equation and other simple models for the growth of a single species, there is an equilibrium at zero population, regardless of the value of the growth rate. Thus, the system (5.2) is assumed to satisfy $f(x, \mu) = 0$ for all μ, so that $x = 0$ is an equilibrium for all parameter values. Since the saddle-node families contain parameter values for which there are *no* equilibria near the point of bifurcation, this situation is qualitatively different, called a *transcritical bifurcation* or *exchange of stability.*

The normal form for a transcritical bifurcation is given by equation (5.2)

$$\dot{x} = \mu\, x - x^2 = x\,(\mu - x)$$

which looks like the logistic equation (see equation (4.3)). Fig. 5.9 shows the function $f(x)$ as μ varies. Note that there is an equilibrium $x = 0$ for *all* values of μ.

From $f(x) = 0$, we have fixed points $x^* = 0$ and $x^* = \mu$. The derivative of $f(x)$ with respect to x is $f'(x) = \mu - 2\,x$. Therefore,

(i) when $\mu < 0$,

$$f'(x^* = 0) = \mu < 0 \quad \Longrightarrow \quad x^* = 0 \text{ is stable,}$$
$$f'(x^* = \mu) = \mu - 2\,\mu = -\,\mu > 0 \quad \Longrightarrow \quad x^* = \mu \text{ is unstable;}$$

(ii) when $\mu = 0$, $x^* = 0$ and $f'(0) = 0$ which is a critical point, and the stability of the fixed point is not determined;

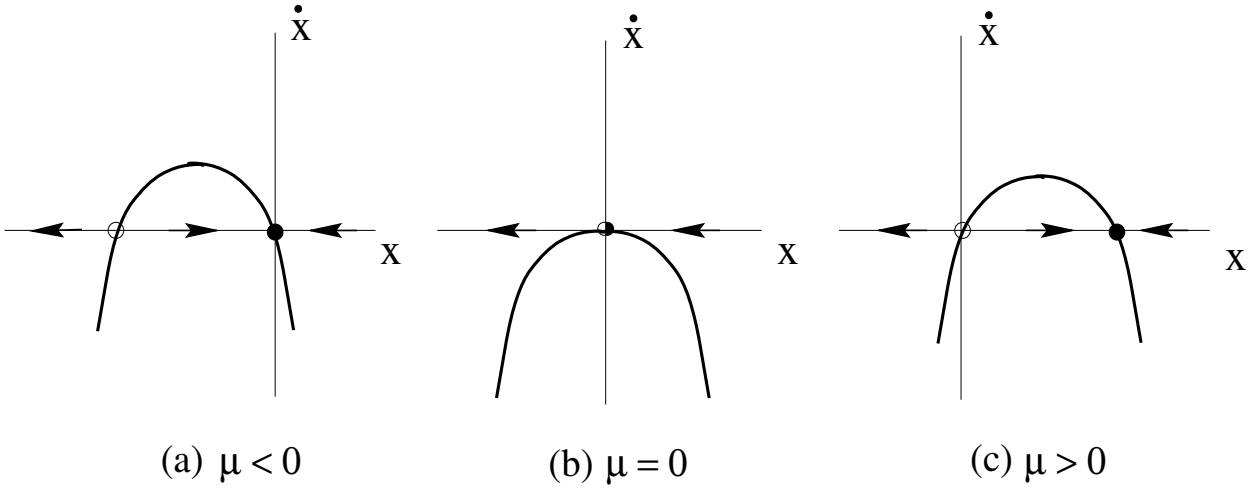


Fig. 5.9 Fixed points of transcritical bifurcation and their stability.

(i) and when $\mu > 0$,

$$f'(x^* = 0) = \mu > 0 \quad \Longrightarrow \quad x^* = 0 \text{ is unstable,}$$
$$f'(x^* = \mu) = \mu - 2\,\mu = -\,\mu < 0 \quad \Longrightarrow \quad x^* = \mu \text{ is stable.}$$

To summarize the above results: for $\mu < 0$, there is an unstable fixed point at $x^* = \mu$ and a stable fixed point at $x^* = 0$. As μ increases, the unstable fixed point approaches the origin, and coalesces with it when $\mu = 0$. Finally, when $\mu > 0$, the origin has become unstable, and $x^* = \mu$ is now stable. This phenomenon is called an *exchange of stabilities* occurring between two fixed points. It should be pointed that as μ increases to cross zero, the number of roots (fixed points) of $f(x)$ does not change, but the stability of the origin changes.

It also should be noted that the important difference between the saddle-node and transcritical bifurcations: in the transcritical case, the two fixed points don't disappear after the bifurcation – instead they just switch their stability. Fig. 5.10 shows the bifurcation diagram for the transcritical bifurcation.

EXAMPLE 5.4. Consider the bifurcation of the following DE

$$\dot{x} = x\,(\mu - e^x) \tag{5.32}$$

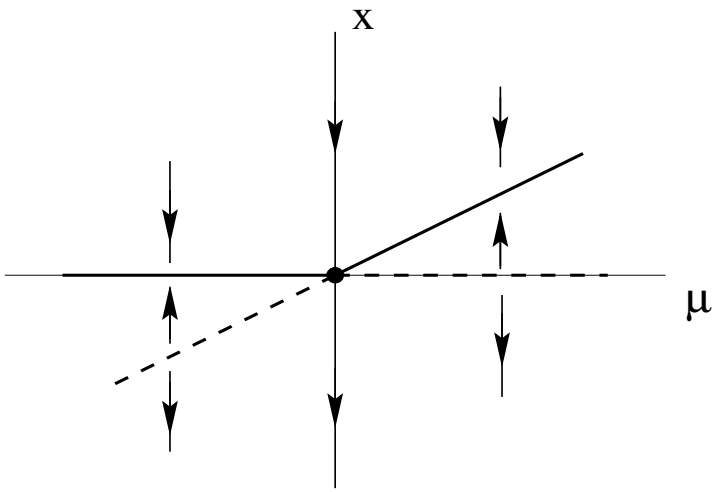


Fig. 5.10 Transcritical bifurcation, —— stable; - - - - unstable.

by using (a) an analytical approach; and (b) a graphical method.

Solution:

(a) Note that $x = 0$ is a fixed point for all values of μ. Another fixed point is given by $e^x = \mu$, so $x = \ln \mu$ if $\mu > 0$. $f'(x) = \mu - e^x (1 + x)$. Thus,

$$f'(0) = \mu - 1 \qquad \text{and} \qquad f'(\ln \mu) = -\mu \ln \mu. \tag{5.33}$$

The critical point may be found from

$$\begin{aligned} f(x) &= x(\mu - e^x) = 0 \\ f'(x) &= \mu - e^x (1 + x) = 0. \end{aligned} \tag{5.34}$$

Solving (5.34) yields the critical point: $\mu_c = 1$ $x_c = 0$. Hence,

$$\text{Fixed point } x^* = 0 \text{ is } \begin{cases} \text{stable} & \text{if } \mu < 1; \\ \text{unstable} & \text{if } \mu > 1. \end{cases} \tag{5.35}$$

And

$$\text{Fixed point } x^* = \ln \mu \text{ is } \begin{cases} \text{unstable} & \text{if } 0 < \mu < 1; \\ \text{stable} & \text{if } \mu > 1. \end{cases} \tag{5.36}$$

The bifurcation diagram is given in Fig. 5.11. Note that $x^* = \ln \mu$ does not exist when $\mu \leq 0$.

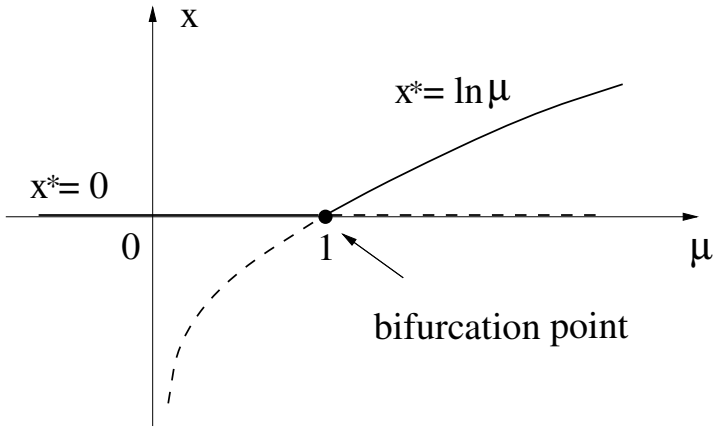


Fig. 5.11 Bifurcation diagram for Example 5.4. —— stable; - - - - unstable.

(b) Equation (5.32) can be rewritten as

$$\dot{x} = x\,(\mu - e^x) = \mu\,x - x\,e^x = f(x) - g(x) \quad \text{with} \quad f(x) = \mu\,x \quad \text{and} \quad g(x) = x\,e^x \tag{5.37}$$

where $f(x)$ is a straight line while $g(x)$ is a simple function. The qualitative behavior of the DE (5.32) is demonstrated in Fig. 5.12.

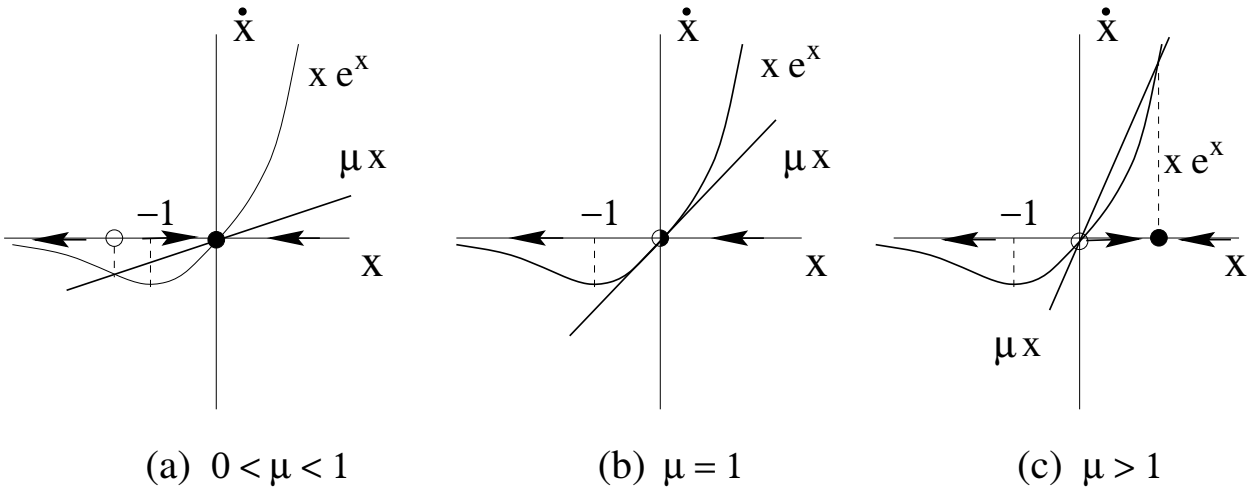


Fig. 5.12 Fixed points and bifurcation for Example 5.4.

EXAMPLE 5.5. Consider the DE

$$\dot{x} = \mu\,x - \ln(1+x). \tag{5.38}$$

Classify the type of bifurcations, and find the critical value of μ and sketch the bifurcation diagram in the $(x,\, \mu)$ plane

Solution: Here $f_\mu(x) = \mu\, x - \ln(1+x)$ for $x > -1$. Fixed points are determined by the $f_\mu(x) = 0$. However, we cannot find explicit expression for this case. We can use graphic approach to consider the fixed points and their stabilities. Rewrite (5.38) as

$$\dot{x} = \mu\, x - \ln(1+x) = f(x) - g(x) \qquad \text{with } f(x) = \mu\, x \quad \text{and} \quad g(x) = \ln(1+x)\,. \tag{5.39}$$

The bifurcation point can be found from the equations

$$\mu\, x = \ln(1+x) \qquad \text{and} \qquad \mu = \frac{1}{1+x} \;\; \text{(i.e., } f\,'(x) = g\,'(x)). \tag{5.40}$$

From the second equation we have $x = \dfrac{1}{\mu} - 1$ and then substitute it into the first equation to obtain

$$1 - \mu + \ln \mu = 0$$

which gives the bifurcation point: $\mu_c = 1$ (and so , $x_c = 0$). The diagrams showing the qualitative behavior of the vector field are given in Fig. 5.13. It is seen that $x = 0$ is an equilibrium for all value of μ. A new equilibrium bifurcates from the critical point and these two equilibria exchange their stability at the critical point. Thus, it is a transcritical bifurcation. The bifurcation diagram is depicted in Fig. 5.14. It should be noted that the nonzero fixed point does not exist if $\mu \le 0$.

Although we cannot find the explicit solution for the nonzero fixed point, we can still find its stability condition analytically. First note that $x_1^* = 0$ is an equilibrium for all value of μ; and the non-trivial equilibrium is defined implicitly as

$$\mu = \frac{\ln(1+x_2^*)}{x_2^*} \qquad \text{for } \; x_2^* > -1 \tag{5.41}$$

where the value of μ at $x_2^* = 0$ can be defined as the limit $\lim\limits_{x_2^* \to 0} \dfrac{\ln(1+x_2^*)}{x_2^*} = 1$. The derivative of f w.r.t. x is

$$f\,'(x) = \mu - \frac{1}{1+x}\,. \tag{5.42}$$

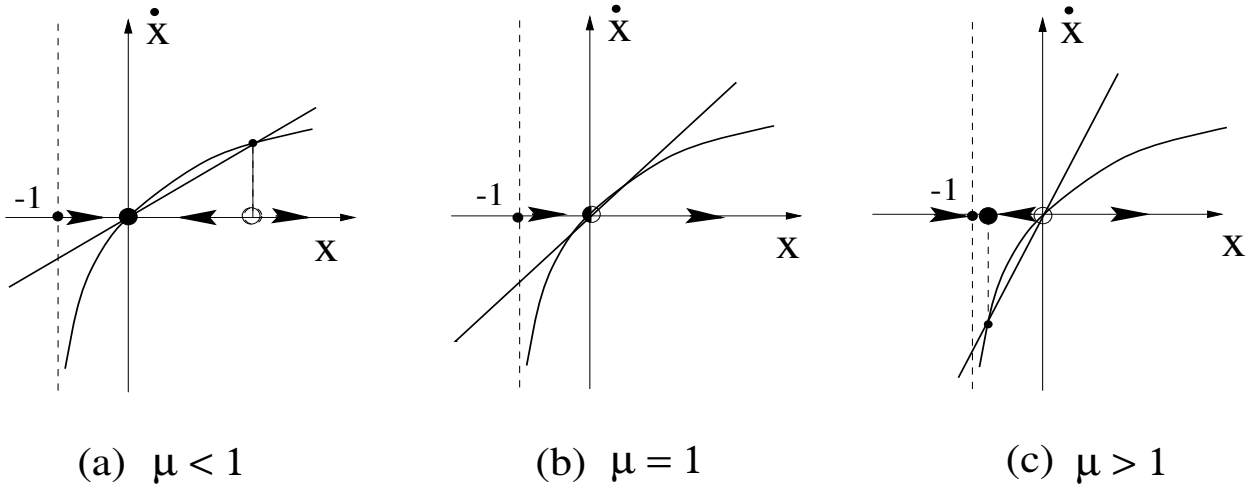


Fig. 5.13 Fixed points of system (5.38).

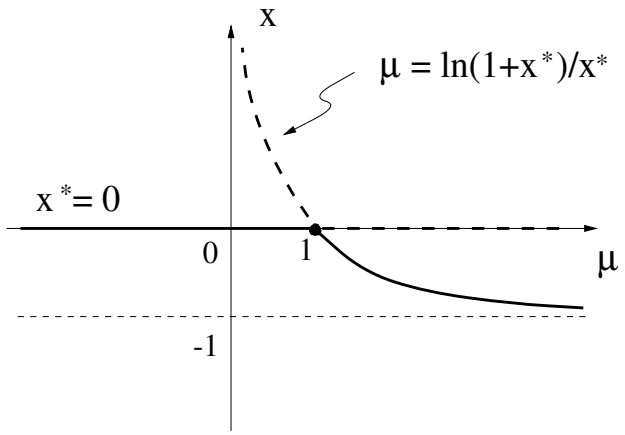


Fig. 5.14 Bifurcation diagram of system (5.38).

Thus, for the equilibrium x_1^*, we have

$$f'(x_1^*) = \mu - 1 \begin{cases} < 0 & \text{if } \mu < 1 \implies x_1^* \text{ stable;} \\ > 0 & \text{if } \mu > 1 \implies x_1^* \text{ unstable;} \\ = 0 & \text{if } \mu = 1 \implies \mu_c = 1 \text{ is a bifurcation point.} \end{cases} \tag{5.43}$$

For the equilibrium x_2^*, the derivative of f w.r.t. x evaluated at x_2^* is

$$f'(x_2^*) = \frac{\ln(1+x_2^*)}{x_2^*} - \frac{1}{1+x_2^*} = (-x_2^*)\frac{\left[\dfrac{x_2^*}{1+x_2^*} - \ln(1+x_2^*)\right]}{(x_2^*)^2} = (-x_2^*)\frac{du}{dx_2^*} \tag{5.44}$$

where $\dfrac{du}{dx_2^*} = -\dfrac{1}{(x_2^*)^2}\left[\ln(1+x_2^*) - \dfrac{x_2^*}{1+x_2^*}\right]$. We now want to prove that $\dfrac{du}{dx_2^*} < 0$ for

$x_2^* \in (-1, \infty)$. In other words, u is a monotonic function of x_2^*, or x_2^* is a monotonic function of μ. To do this, let

$$g(x_2^*) = \ln(1 + x_2^*)\,, \tag{5.45}$$

and suppose $-1 < x_2^* < 0$, then

$$\ln(1 + x_2^*) = g(x_2^*) - g(0) = g'(\xi)\, x_2^* = \frac{x_2^*}{1+\xi} \qquad \text{for some } \xi \in (x_2^*,\, 0) \tag{5.46}$$

by the Lagrange theorem. Now, from $-1 < x_2^* < \xi < 0$, we have

$$\begin{aligned} -1 < x_2^* < \xi < 0 \iff & \; 1 + \xi > 1 + x_2^* \\ \iff & \; \frac{1}{1+\xi} < \frac{1}{1+x_2^*} \\ \iff & \; \frac{x_2^*}{1+\xi} > \frac{x_2^*}{1+x_2^*} \qquad (\text{since } x_2^* < 0) \\ \iff & \; \ln(1 + x_2^*) > \frac{x_2^*}{1+x_2^*}\,. \end{aligned} \tag{5.47}$$

Thus, $\dfrac{d\mu}{dx_2^*} < 0$ for $x_2^* < 0$. Similarly, we can prove that $\dfrac{d\mu}{dx_2^*} < 0$ for $x_2^* > 0$. Hence,

$$\frac{d\mu}{dx_2^*} < 0 \quad \left(\frac{dx_2^*}{d\mu} < 0\right) \qquad \text{for } x_2^* > -1 \tag{5.48}$$

which indicates that x_2^* is a monotonic function of μ. Also we know that when $x_2^* = 0$, $\mu = 1$. So, $x_2^* < 0 \iff \mu > 1$ and $x_2^* > 0 \iff \mu < 1$. Moreover,

$$f'(x_2^*) = (-x_2^*)\frac{d\mu}{dx_2^*} \begin{cases} > 0 & \text{if } \mu < 1 \; (x_2^* > 0) \implies x_2^* \text{ unstable;} \\ < 0 & \text{if } \mu > 1 \; (x_2^* < 0) \implies x_2^* \text{ stable;} \\ = 0 & \text{if } \mu = 1 \; (x_2^* = 0) \implies \mu_c = 1 \text{ is a bifurcation point.} \end{cases} \tag{5.49}$$

This shows that the two equilibria x_1^* and x_2^* exchange stability at the critical point $\mu_c = 1$. It is a transcritical bifurcation.

5.3 Pitchfork Bifurcation

A second setting in which the saddle-node does not occur involves systems which have a symmetry. Many physical problems are formulated so that the equations defining the system do have symmetries of some kind. For example, the Duffing equation is invariant

under the transformation $(x, y) \to (-x, -y)$ and the Lorenz equation is invariant under the transformation $(x, y, z) \to (-x, -y, z)$. In the buckling example of Fig. 4.1, the beam is stable in the vertical position if the load is small. In this case there is a stable fixed point corresponding to zero deflection. But if the load exceeds the buckling threshold, the beam may buckle to *either* the left or the right. The vertical position has gone unstable, and two new symmetrical fixed points, corresponding to left- and right-buckled configurations, have been born. In one dimension, a differential equation (5.3)

$$\dot{x} = \mu x - x^3$$

is symmetric or *equivalent* with respect to the symmetry $x \to -x$ if $f(-x; \mu) = -f(x; \mu)$. Thus, the equivalent vector fields are ones for which $f(x; \mu)$ is an odd function of x. In particular, all such equations have an equilibrium at 0. The transcritical bifurcation cannot occur in these systems, however, because an odd function $f(x; \mu)$ cannot satisfy the condition $\dfrac{\partial^2 f(x; \mu)}{\partial x^2} \neq 0$ required by the transcritical bifurcation. If this condition is replaced by the transversality condition $\dfrac{\partial^3 f(x; \mu)}{\partial x^3} \neq 0$, then one obtains the *pitchfork bifurcation.* At the point of bifurcation, the stability of the trivial equilibrium changes, and a new *pair* of equilibria (related by the symmetry) appear to one side of the point of bifurcation in parameter space, as in Fig. 5. 15.

This bifurcation diagram looks much like the period-doubling bifurcation (see Fig. 4.13). *But here these are all fixed points, not cycles.* It is noted that the direction of the bifurcation and the stability of the branches is determined by the sign of $\dfrac{\partial^2 f(x; \mu)}{\partial x^2}$ or $\dfrac{\partial^3 f(x; \mu)}{\partial x^3}$. In the last case, if $\dfrac{\partial^2 f(x; \mu)}{\partial x^2}$ is negative, then branches occur "above" the bifurcation value and it is called a *supercritical* pitchfork bifurcation (as shown in Fig. 5.15), otherwise it is called a *subcritical* bifurcation.

The pitchfork bifurcation can also be illustrated in vector field $f(x)$. Consider equation (5.3) and note that it is invariant under the change of variables $x \to -x$. Fig. 5.16 shows the vector field for different values of μ. When $\mu < 0$, the origin is the only fixed point, and it is stable. When $\mu = 0$ the origin is still stable, but much more weakly so, since

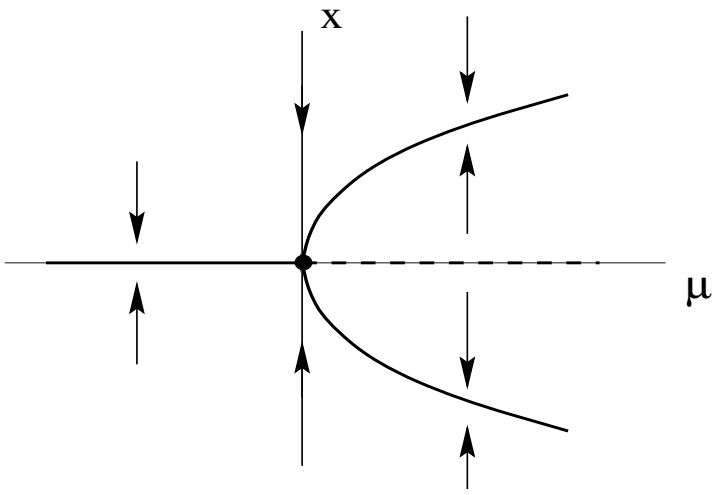


Fig. 5.15 Supercritical pitchfork bifurcation, —— stable; - - - - unstable.

the linearization vanishes. Now solutions no longer decay exponentially fast – instead the decay is a much slower algebraic function of time (the solution of the D.E. $\dot{x} = -x^3$ is $x(t) = [\,2\,(t+C)\,]^{-1/2}$). This lethargic decay is called *critical slowing down* in the physics literature. Finally, when $\mu > 0$, the origin has become unstable. Two new stable fixed points appear on either side of the origin, symmetrically located at $x^* = \pm\sqrt{\mu}$. The reason for the term "pitchfork" can be seen from Fig. 5.15. Actually, pitchfork trification might be a better word! Pitchfork bifurcation is also called *flip bifurcation.*

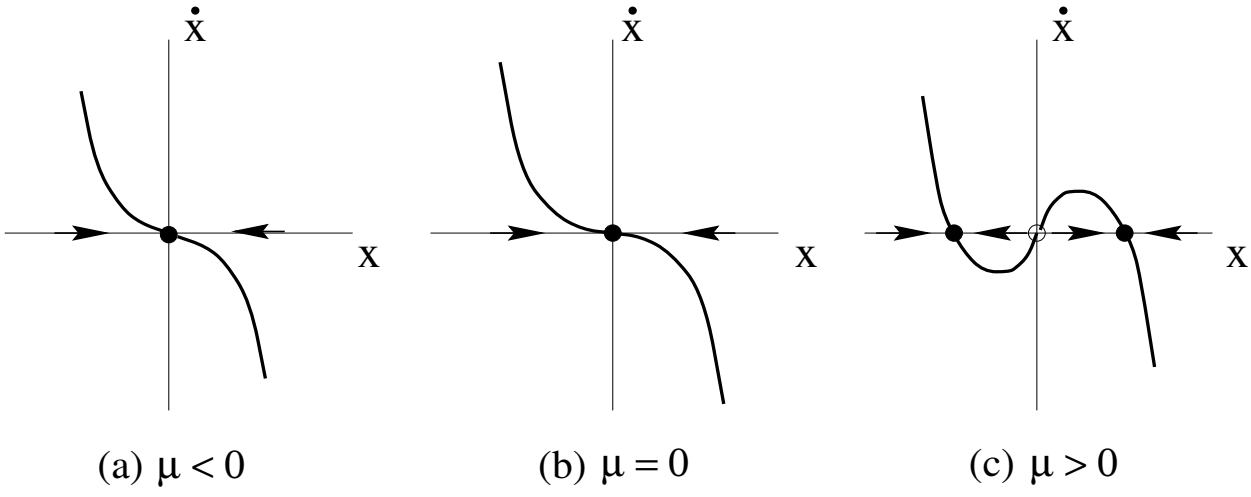


Fig. 5.16 Fixed points of pitchfork bifurcation and their stability.

EXAMPLE 5.6. Equations similar to

$$\dot{x} = -x + \beta \tanh x \tag{5.50}$$

arise in statistical mechanical models of magnets and neural networks. Show that this equation undergoes a supercritical pitchfork bifurcation as β is varied. Then give a *numerically accurate* plot of the fixed points for each β.

Solution: We use the strategy of Example 5.2 to find the fixed points. The graphs of $y = x$ and $y = \beta \tanh x$ are shown in Fig. 5.17; their intersections correspond to fixed points. The key thing to realize is that as β increases, the tanh curve becomes steeper at the origin (its slope there is β). Hence for $\beta < 1$ the origin is the only fixed point. A pitchfork bifurcation occurs at $\beta = 1$, $x^* = 0$, when the tanh curve develops a slope of 1 at the origin. Finally, when $\beta > 1$, two new stable fixed points appear, and the origin becomes unstable.

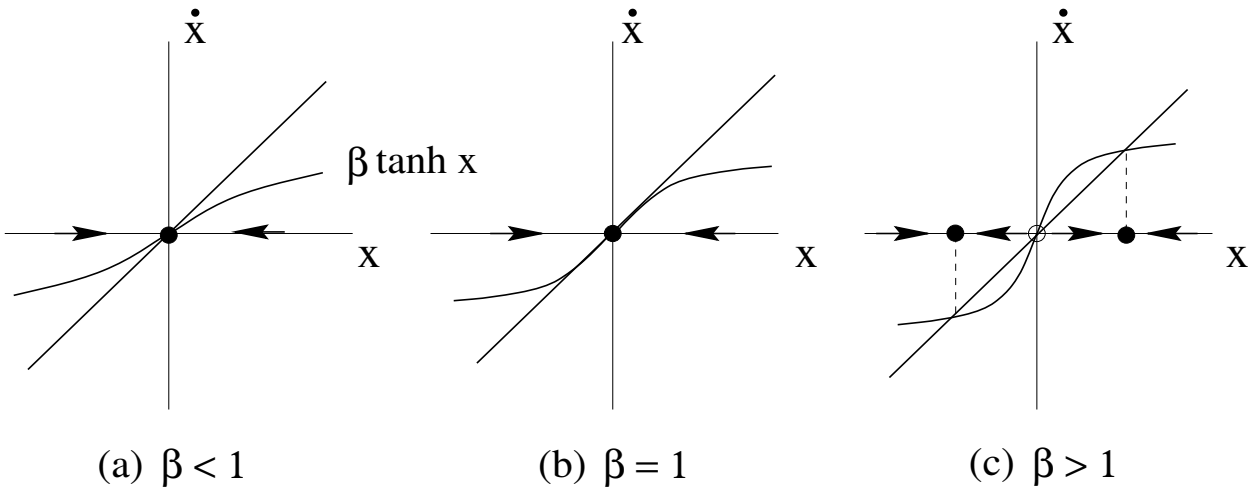


Fig. 5.17 Fixed points of Example 5.6.

To compute the fixed point x^* for each β, we only need to look for the nontrivial fixed point (the trivial one is $x^* = 0$). One approach is to solve the equation $x^* = \beta \tanh x^*$ numerically, using the Newton-Raphson method or some other root-finding scheme. But there is an easier way, which comes from changing our point of view. Instead of studying the dependence of x^* on β, we think of x^* as the *independent* variable, and then compute $\beta = \dfrac{x^*}{\tanh x^*}$. This gives us a table of pairs of (x^*, β). For each pair, we plot β horizontally

and x^* vertically. This yields the bifurcation diagram (Fig. 5.18). The shortcut used here exploits the fact that $f(x,\,\beta) = -\,x + \beta\,\tanh x$ depends more simply on β than on x. This is frequently the case in bifurcation problems – the dependence on the control parameter is usually simpler than the dependence on x.

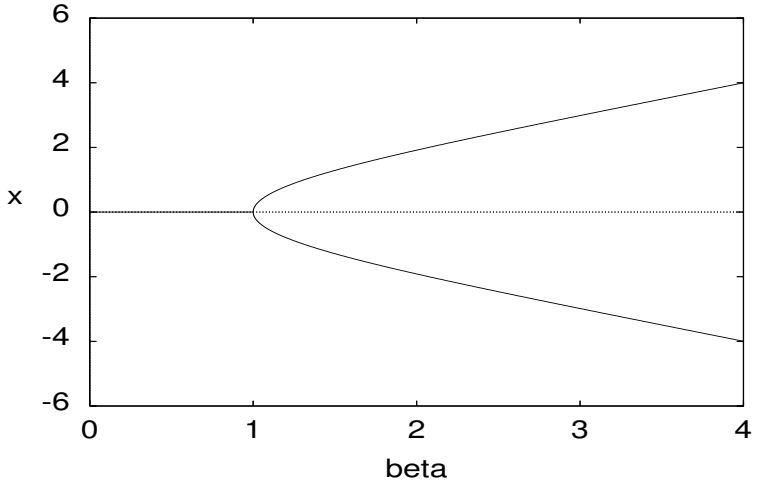


Fig. 5.18 Pitchfork bifurcation of Example 5.6.

EXAMPLE 5.7. Consider the DE

$$\dot{x} = x + \frac{\mu\,x}{1+x^2}. \tag{5.51}$$

Classify the type of bifurcations, and find the critical value of μ and sketch the bifurcation diagram in the $(x,\,\mu)$ plane

Solution: The fixed points are given by

$$x_1^* = 0 \qquad \text{and} \qquad x_2^* = \pm\sqrt{-\,(1+\mu)}\,, \quad (\mu < -\,1)\,. \tag{5.52}$$

It is a pitchfork bifurcation. To find the stability of these fixed points, by using

$$f\,'(x) = 1 + \frac{\mu\,(1-x^2)}{(1+x^2)^2} \tag{5.53}$$

one obtains

$$f\,'(x_1^*) = 1+\mu \begin{cases} < 0 \;\text{ if }\; \mu < -\,1 \quad \Longrightarrow \quad x_1^* = 0 \;\text{ is stable;} \\ > 0 \;\text{ if }\; \mu > -\,1 \quad \Longrightarrow \quad x_1^* = 0 \;\text{ is unstable;} \end{cases} \tag{5.54}$$

and the bifurcation point is $\mu_c = -1$. Similarly, for the fixed point x_2^*, we find

$$f'(x_2^*) = 1 + \frac{\mu\,[\,1-(x_2^*)^2\,]}{[\,1+(x_2^*)^2\,]^2} = 1 - \frac{[\,1+(x_2^*)^2\,][\,1-(x_2^*)^2\,]}{[\,1+(x_2^*)^2\,]^2} = \frac{2\,(x_2^*)^2}{1+(x_2^*)^2} > 0\,, \qquad (5.55)$$

thus, the nonzero fixed point is unstable, implying that this is a subcritical pitchfork bifurcation (Fig. 5.19).

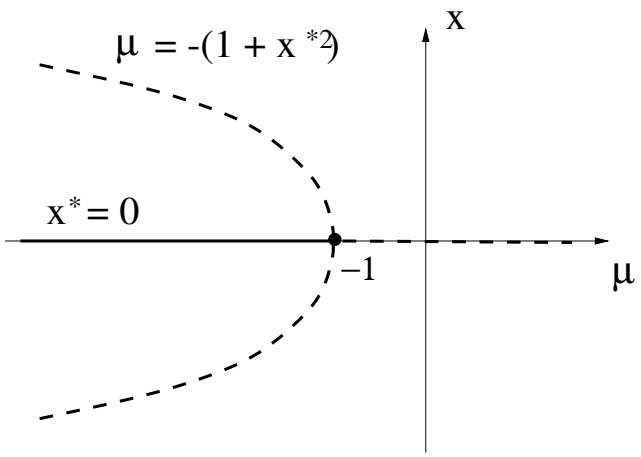


Fig. 5.19 Bifurcation diagram of system (5.51).

In the *supercritical* pitchfork bifurcation discussed above, the cubic term is *stabilizing*: it acts as a restoring force that pulls $x(t)$ back toward $x = 0$. If instead the cubic term were *destabilizing*, as in

$$\dot{x} = x\,(\mu + x^2)\,, \qquad (5.56)$$

then it is called *subcritical* pitchfork bifurcation. Fig. 5.20 shows the bifurcation diagram. Compared to Fig. 5.15, the pitchfork is inverted. The nonzero fixed points $x^* = \pm\sqrt{\mu}$ are *unstable*, and exist only *below* the bifurcation ($\mu < 0$), which motivates the term "subcritical." More importantly, the origin is stable for $\mu < 0$ and unstable for $\mu > 0$, as in the supercritical case, but now the instability for $\mu > 0$ is not opposed by the cubic term – in fact the cubic term lends a helping hand in driving the trajectories out to infinity! This effect leads to *blow-up*: one can show that $x(t) \to \pm\infty$ in finite time, starting from an initial condition $x_0 \neq 0$.

In real physical systems, such an explosive instability is usually opposed by the stabilizing influence of higher-order terms. Assuming that the systems is still symmetric under $x \to -x$,

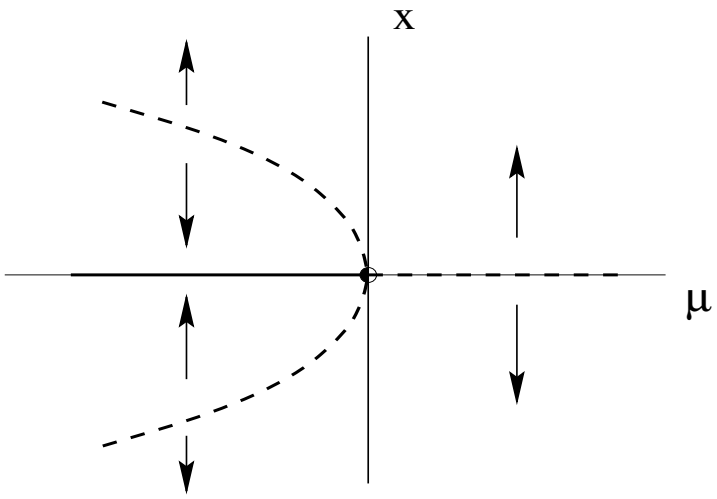


Fig. 5.20 Subcritical pitchfork bifurcation, —— stable; - - - - unstable.

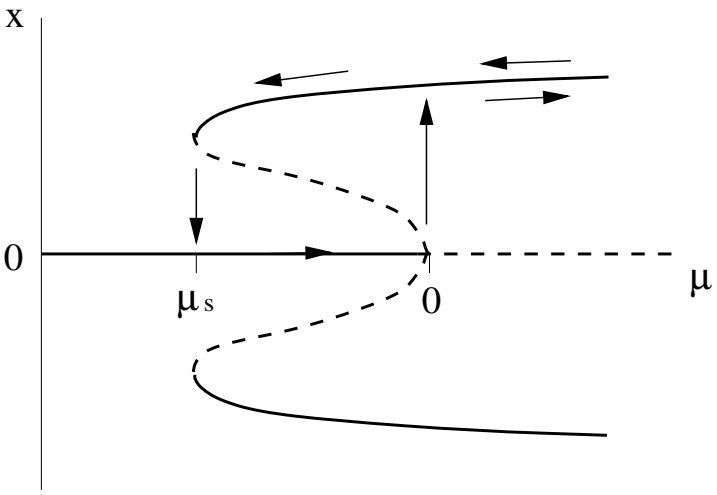


Fig. 5.21 Bifurcation diagram for equation $\dot{x} = \mu\, x + x^3 - x^5$, —— stable; - - - - unstable.

the first stabilizing term must be x^5. Thus, the canonical example of a system with a subcritical pitchfork bifurcation is

$$\dot{x} = \mu\, x + x^3 - x^5. \tag{5.57}$$

There is no loss in generality in assuming that the coefficients of x^3 and x^5 are 1. The detailed analysis is left to you. But the main results are summarized here. Fig. 5.21 illustrates the bifurcation diagram for (5.57). For small x, the picture looks like Fig. 5.20:

the origin is locally stable for $\mu < 0$, and two backward-bending branches of unstable fixed points bifurcate from the origin when $\mu = 0$. The new feature, due to the x^5 term, is that the unstable branches turn around and become stable at $\mu = \mu_s$, where $\mu_s < 0$. These stable **large amplitude** branches exist for all $\mu > \mu_s$.

1) In the range $\mu_s < \mu < 0$, two qualitatively different stable states coexist, namely the origin and the large-amplitude fixed points. The initial condition x_0 determines which fixed point is approached as $t \to \infty$. One consequence is that the origin is stable to small perturbations, but not to large ones – in this sense the origin is *locally* stable, but not *globally* stable.

2) The existence of different stable states allows for the possibility of **jumps** and **hysteresis** as μ is varied. Suppose we start the system in the state $x^* = 0$, and then slowly increase the parameter μ (indicated by an arrow along the μ-axis of Fig. 5.21). Then the state remains at the origin until $\mu = 0$, when the origin loses stability. Now the slightest change will cause the state to *jump* to one of the large-amplitude branches. With further increases of μ, the state moves out along the large-amplitude branch. If μ is now decreased, the state remains on the large-amplitude branch, even when μ is decreased below 0! We have to lower μ even further (down pass μ_s) to get the state to jump back to the origin. This lack of reversibility as parameter is varied is called *hysteresis*.

3) The bifurcation at μ_s is a saddle-node bifurcation, in which stable and unstable fixed points are born as μ increased.

5.4 Oscillations: Hopf Bifurcation

There is a difference between discrete maps and continuous differential equations. We had oscillations in the logistic map in one (parameter) dimension. For a continuous DE, we need, at least, 2 dimension to have oscillations. Something like the period doubling bifurcation may happen then in the DE case. Oscillation coming from a fixed point is described by the *Hopf bifurcation*. This corresponds to eigenvalues crossing the imaginary axis

from negative to positive real parts (Fig. 5.22). More frequencies become very complicated, mathematically speaking.

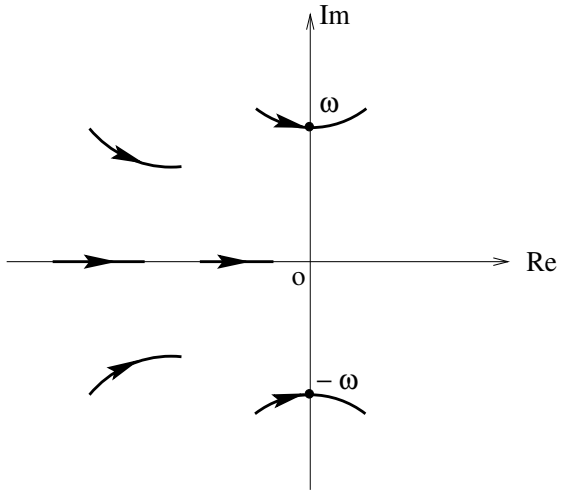


Fig. 5.22 Transversality of Hopf bifurcation.

A clue to what happens in the generic bifurcation problem involving an equilibrium with pure imaginary eigenvalues can be gained from examining linear systems in which there is a change of this type. For example, consider the system

$$\begin{aligned} \dot{x} &= \mu\, x - \omega\, y \\ \dot{y} &= \omega\, x + \mu\, y \end{aligned} \tag{5.58}$$

whose solutions have the form

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = e^{\mu t} \begin{bmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{bmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \tag{5.59}$$

When $\mu < 0$, solutions spiral into the origin, and when $\mu > 0$, solutions spiral away from the origin. When $\mu = 0$, all solutions are periodic. Even in a one-parameter family of equations, it is highly special to find a parameter value at which there is a whole family of periodic orbits, but there is still a surface of periodic orbits which appears in the general problem.

The normal form theorem gives us the required information about how the generic problem differs from the system (5.58). By smooth changes of coordinates, the Taylor series

of degree 3 for the general problem can be brought to the following form

$$\begin{aligned} \dot{x} &= \left[d\,\mu + a\,(x^2+y^2) \right] x - \left[\omega + c\,\mu + b\,(x^2+y^2) \right] y \\ \dot{y} &= \left[\omega + c\,\mu + b\,(x^2+y^2) \right] x + \left[d\,\mu + a\,(x^2+y^2) \right] y \end{aligned} \tag{5.60}$$

which is expressed in polar coordinates as

$$\begin{aligned} \dot{r} &= \left(d\,\mu + a\,r^2 \right) r \\ \dot{\theta} &= \omega + c\,\mu + b\,r^2 \,. \end{aligned} \tag{5.61}$$

where $r = \sqrt{x^2+y^2}$ and $\theta = \tan^{-1}(y/x)$. Since the $\dot{r}$ equation in (5.61) separates form θ, we see that there are periodic orbits of (5.60) which are circles $r = \text{const.}$, obtained from the nonzero solutions of $\dot{r} = 0$ in (5.61). If $a \neq 0$ and $d \neq 0$ these solutions lie along the parabola $\mu = -\left(\frac{a}{d}\right) r^2$. This relation is shown in Fig. 5.23. The bifurcating periodic solutions depicted in three dimensional space is given in Fig. 5.24.

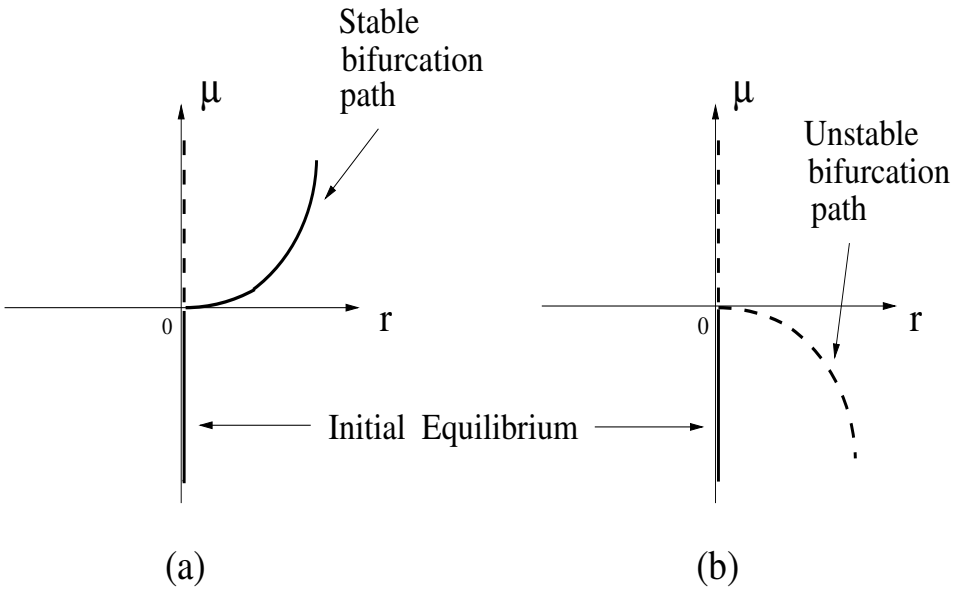


Fig. 5.23 Post-critical bifurcation path for a Hopf bifurcation: (a) stable; (b) unstable.

This production of cycle bifurcations opens the door for a development of chaos in continuous dynamical systems paralleling to the period-doubling cascade of the discrete

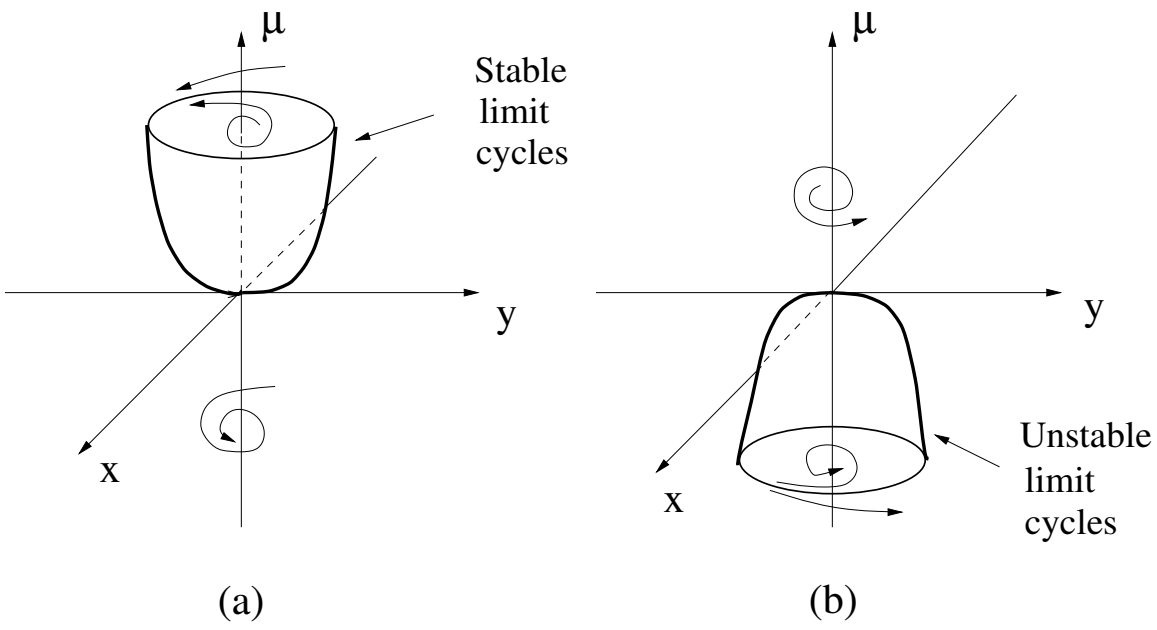


Fig. 5.24 Bifurcating periodic solutions (limit cycles): (a) stable; (b) unstable.

map. Indeed, this is seen in continuous systems of differential equations. It also leads to failures of the traditional analytical approaches as we shall see in the following discussion of perturbation theory.

Chapter 6

PERTURBATION THEORY

This chapter is to present several useful and often used perturbation methods in studying nonlinear dynamical systems: Lindstedt-Poincaré procedure, time averaging and multiple scales (also called multiple time scales).

6.1 Introduction

Most of the physical problems facing engineers, physicists, and applied mathematicians today exhibit certain essential features which preclude exact analytical solutions. Some of these features are nonlinearities, variable coefficients, complex boundary shapes, and nonlinear boundary conditions at known or, in some cases, unknown boundaries. Even if the exact solution of a problem can be found explicitly, it may be useless for mathematical and physical interpretation or numerical evaluation. Examples of such problems are Bessel functions of large argument and large-order and doubly periodic functions. Thus, in order to obtain information about solutions of equations, we are forced to resort to approximations, numerical solutions, or combinations of both. Foremost among the approximation methods are perturbation (asymptotic) methods.

Perturbation methods arisen from the study of celestial mechanics in the end of 19th century (Poincaré) and the work of many authors (e.g. Lindstedt). Perturbation theory can be a very messy business: Delauney in the mid 19th century derived a seventh order perturbation approximation to the moon's orbit. (Recall that the two-body problem is integrable but it is not possible to obtain closed form solutions for the three-body problem.) He obtained an equation which covered 170 printed pages. Another example is from the area of kinetic theory. If one attempts to explain fluid mechanics, in particular, Navier-Stokes theory by deriving them from first principles, one can use perturbation techniques to give higher order approximations, giving rise the *Burnett stress tensor* (Burnett equation),

which is very much complicated. When direct methods like those used to solve the logistic equation, cannot be found and do not exist, then perturbation methods are our last resort for direct analytical techniques short of going to a computer.

According to perturbation techniques, the solution is represented by the first few terms of an *asymptotic expansion*, usually not more than three terms. The expansion may be carried out in terms of a parameter (small or large) which appears naturally in the equations, or which may be artificially introduced for convenience. Such expansions are called *parameter perturbations*. Alternatively, the expansions may be carried out in terms of a coordinate (either small or large); these are called *coordinate perturbations*. Note that perturbations to large parameters (or coordinates) can be transformed to perturbations to small parameters (or coordinates). The difficulty other than complexity is that perturbation methods are ultimately series solutions in a small parameter. Series solutions have often a finite radius of convergence. This suggests that we will not be able to get all solutions of a nonlinear differential equation by virtue of a fixed reference solutions to expand about. A more subtle problem arises if one thinks of our experience with "frequencies" in the transition to chaos. In perturbation theory one has to treat oscillations carefully. A single frequency or many distinct frequencies can be treated. However, in the transition to chaos the numbers of frequencies become infinite and densely distributed. This ultimately leads to a general failure of perturbation methods.

Suppose we have a system described by

$$\ddot{x} + x + \epsilon\,(x^2 - 1)\,\dot{x} = 0 \qquad \text{(Van der Pol equation)} \tag{6.1}$$

or

$$\ddot{x} + x + \epsilon\,x^3 = 0 \qquad \text{(Duffing equation)} \tag{6.2}$$

where ϵ is a small non-negative real number (i.e., $0 \leq \epsilon \ll 1$). This kind of systems are called *weakly nonlinear systems* and perturbation methods can be applied to find approximate solutions. Since we can think in this way that the above two nonlinear equations are obtained from the linear DE

$$\ddot{x} + x = 0 \tag{6.3}$$

by a small perturbation (ϵ terms although they are nonlinear), so it's reasonable to believe that the solutions of the nonlinear DE can be obtained by perturbing the solutions of the linear DE. This leads to a technique – perturbation.

Let's consider a simple physical example – the real pendulum moving in a vertical plane, shown in Fig. 6.1. It is assumed that the mass of the rod supporting the bob is neglected. (Also, frictional force is neglected.)

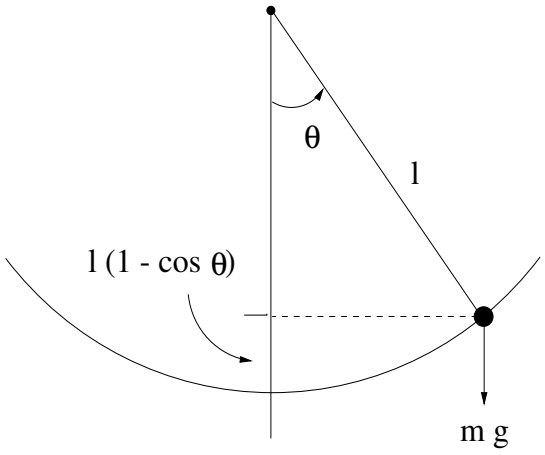


Fig. 6.1 Pendulum.

Let l be the (constant) length of the rod. The bob of the pendulum moves along a circle of radius l. If $\theta(t)$ is the counterclockwise angle from the vertical to the rod at time t, then the angular velocity of the bob is $\dfrac{d\theta}{dt}$ and the velocity is $l\,\dfrac{d\theta}{dt}$. The downward gravitational force $m\,g$, g being the gravitational constant, has component $-\,m\,g\,sin\,\theta$ tangent to the circle; this is the force on the bob that produces motion. Therefore, the total force tangent to the circle at time t is

$$F = -\,m\,g\,\sin\theta. \tag{6.4}$$

From Newton's law $F = a\,m$, where $a = l\,\dfrac{d^2\theta}{dt^2}$ is acceleration, we have

$$l\,\ddot{\theta} = -\,g\,\sin\theta$$

or

$$\ddot{\theta} + \frac{g}{l}\,\sin\theta = 0. \tag{6.5}$$

Let $\omega_0 = \sqrt{\frac{g}{l}}$, which is natural angular velocity, then (6.5) becomes

$$\ddot{\theta} + \omega_0^2 \sin\theta = 0. \tag{6.6}$$

$\sin\theta$ can be expressed in Taylor expansion:

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots \tag{6.7}$$

which converges for $\theta \in (-\infty, \infty)$. For small θ, $\sin\theta \approx \theta$, thus we have the linearized system

$$\ddot{\theta} + \omega_0^2 \theta = 0 \tag{6.8}$$

and can solve this system exactly to obtain solution

$$\theta(t) = A \cos\omega_0 t + B \sin\omega_0 t \tag{6.9}$$

where A and B are constants determined from the initial conditions. This equation represents periodic solutions with an identical frequency ω_0 (the natural frequency) no matter what the amplitude of the motion may be. Is this true? Why do we stop at the linear term θ? If we take one nonlinear term, i.e., up to the third order term, then the DE (6.6) is approximated by

$$\ddot{\theta} + \omega_0^2 \left(\theta - \frac{\theta^3}{6}\right) = 0 \tag{6.10}$$

which can be transformed to

$$\ddot{x} + \omega_0^2 x - \frac{1}{6}\epsilon\,\omega_0^2 x^3 = 0 \tag{6.11}$$

by the transformation

$$\theta = \epsilon^{1/2} x \qquad (\epsilon > 0) \tag{6.12}$$

The above non-standard (dimensional) form DE (6.11) will be used to show how to treat a general nonlinear system using a perturbation technique. It is noted that a further transformation can change system (6.11) to a dimensionless one with frequency 1. However, in general cases, it does not simplify the analysis. When $\epsilon = 0$, the nonlinear DE (6.11) becomes the linear DE (6.8) and its solution is give by (6.9).

6.2 Regular Perturbation Procedure

As a first approach, we seek solutions of (6.11) in the form of a power series in ϵ. Thus if $x(t,\epsilon)$ is a solution, we expand it as

$$x(t,\epsilon) = x_0(t) + \epsilon\, x_1(t) + \epsilon^2\, x_2(t) + \ldots \tag{6.13}$$

where the unknown functions $x_k(t)$ are to be determined from the DE (6.11) and the initial conditions. The hope is that all the information is captured by the first few terms – ideally, the first two – and that the higher-order terms represent only tiny corrections. This technique is called *regular perturbation theory* or *direct perturbation method*. It works well on certain classes of problems, but as we'll see, it turns into trouble here.

Differentiating (6.13) with respect to t twice yields

$$\ddot{x} = \ddot{x}_0 + \epsilon\, \ddot{x}_1 + \epsilon^2\, \ddot{x}_2 + \ldots \tag{6.14}$$

and

$$\epsilon\, x^3 = \epsilon\,(x_0 + \epsilon\, x_1 + \epsilon^2\, x_2 + \ldots)^3 = \epsilon\, x_0^3 + 3\,\epsilon^2\, x_0^2\, x_1 + \ldots \tag{6.15}$$

where t is omitted for simplicity. Substituting the above two expressions to (6.11) results in

$$\ddot{x}_0 + \epsilon\, \ddot{x}_1 + \epsilon^2\, \ddot{x}_2 + \ldots + \omega_0^2\,(x_0 + \epsilon\, x_1 + \epsilon^2\, x_2 + \ldots) - \frac{1}{6}\,\omega_0^2\,(\epsilon\, x_0^3 + 3\,\epsilon^2\, x_0^2\, x_1 + \ldots) = 0 \tag{6.16}$$

which can be grouped according to the powers of ϵ as

$$(\ddot{x}_0 + \omega_0^2\, x_0) + \epsilon\,(\ddot{x}_1 + \omega_0^2\, x_1 - \frac{1}{6}\,\omega_0^2\, x_0^3) + \epsilon^2\,(\ddot{x}_2 + \omega_0^2\, x_2 - \frac{1}{2}\,\omega_0^2\, x_0^2\, x_1) + \ldots = 0\,. \tag{6.17}$$

Because this equation should be satisfied for all sufficiently small ϵ, so the coefficients of all powers of ϵ must be vanish separately. Thus, balancing the like powers of ϵ yields a set of differential equations:

$$\epsilon^0 \;:\quad \ddot{x}_0 + \omega_0^2\, x_0 = 0 \qquad \text{(This equation is indeed the linearized DE (6.9))} \tag{6.18}$$

$$\epsilon^1 \;:\quad \ddot{x}_1 + \omega_0^2\, x_1 = \frac{1}{6}\,\omega_0^2\, x_0^3 \tag{6.19}$$

$$\epsilon^2 \;:\quad \ddot{x}_2 + \omega_0^2\, x_2 = \frac{1}{2}\,\omega_0^2\, x_0^2\, x_1 \tag{6.20}$$

etc.

It can be seen from the above equations that the LHS (Left-Hand-Side) of all the equations are identical except the subscripts, while the RHS (Right-Hand-Side) contain variables which have been solved from previous equations, except the ϵ^0-order equation. Therefore, solving an ϵ^k-order perturbation equation depends upon the previous k DEs only. Also, it should be noted that all the equations are linear DEs.

Before we solve the set of linear DEs, we need to find the initial conditions of these DEs. Suppose the initial conditions of the original nonlinear DE (6.11) are given by

$$x(0) = C_1 \qquad \text{and} \qquad \dot{x}(0) = \omega_0 C_2 \tag{6.21}$$

then from equations $x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + \ldots$ and $\dot{x} = \dot{x}_0 + \epsilon \dot{x}_1 + \epsilon^2 \dot{x}_2 + \ldots$, one has

$$\begin{aligned} x(0) &= x_0(0) + \epsilon x_1(0) + \epsilon^2 x_2(0) + \ldots \\ \dot{x}(0) &= \dot{x}_0(0) + \epsilon \dot{x}_1(0) + \epsilon^2 \dot{x}_2(0) + \ldots \end{aligned} \tag{6.22}$$

which, in turn, by balancing the like powers of ϵ, results in

$$x_0(0) = C_1 \ , \qquad \dot{x}_0(0) = \omega_0 C_2 \tag{6.23}$$

and

$$x_n(0) = 0 \ , \qquad \dot{x}_n(0) = 0 \qquad \text{for } n \geq 1 \, . \tag{6.24}$$

Now we solve the initial-value problems one by one. First, from the ϵ^0-order equation (6.18), we find the solution, given by $x_0(t) = A_0 \cos\omega_0 t + B_0 \sin\omega_0 t$, which becomes

$$x_0(t) = C_1 \cos\omega_0 t + C_2 \sin\omega_0 t \tag{6.25}$$

by imposing the initial conditions (6.23).

Next, from the ϵ^1-order equation (6.19), one obtains

$$\begin{aligned} \ddot{x}_1 + \omega_0^2 x_1 &= \frac{1}{6}\omega_0^2 (C_1 \cos\omega_0 t + C_2 \sin\omega_0 t)^3 \\ &= \frac{1}{8}\omega_0^2 (C_1^2 + C_2^2)(C_1 \cos\omega_0 t + C_2 \sin\omega_0 t) \\ &\quad + \frac{1}{24}\omega_0^2 (C_1^2 - 3C_2^2) C_1 \cos 3\omega_0 t + \frac{1}{24}\omega_0^2 (3C_1^2 - C_2^2) C_2 \sin 3\omega_0 t \end{aligned} \tag{6.26}$$

where trigonometric identities have been used. The solutions of the above linear DE includes two parts: a general solution and a particular solution. The general solution, solved from the corresponding *homogeneous* DE

$$\ddot{x}_1 + \omega_0^2\, x_1 = 0 \tag{6.27}$$

is

$$x_1^h = A_1\, \cos\omega_0 t + B_1\, \sin\omega_0 t \tag{6.28}$$

and the particular solution is solved from the corresponding non-homogeneous DE

$$\begin{aligned} \ddot{x}_1 + \omega_0^2\, x_1 \;\; &= \;\; \frac{1}{8}\,\omega_0^2\,(C_1^2 + C_2^2)\,(C_1\, \cos\omega_0 t + C_2\, \sin\omega_0 t) \\ &\quad + \frac{1}{24}\,\omega_0^2\,(C_1^2 - 3\,C_2^2)\,C_1\, \cos 3\omega_0 t + \frac{1}{24}\,\omega_0^2\,(3\,C_1^2 - C_2^2)\,C_2\, \sin 3\omega_0 t \end{aligned} \tag{6.29}$$

which includes two parts: one is contributed from the *resonant forcing* term $\frac{1}{8}\,\omega_0^2\,(C_1^2 + C_2^2)\,(C_1\, \cos\omega_0 t + C_2\, \sin\omega_0 t)$ since the natural frequency is same as the external frequency (ω_0), and the other is produced from the non-resonant forcing term $\frac{1}{24}\,\omega_0^2\,(C_1^2 - 3\,C_2^2)\,C_1\, \cos 3\omega_0 t + \frac{1}{24}\,\omega_0^2\,(3\,C_1^2 - C_2^2)\,C_2\, \sin 3\omega_0 t$. The solutions of these two parts can be determined by using the method of *undetermined coefficients*. For example, the solution for the resonant part can be assumed as

$$x_1^{p_1} = t\,(A\, \cos\omega_0 t + B\, \sin\omega_0 t)\,, \tag{6.30}$$

$$\ddot{x}_1^{p_1} = 2\,(-\,\omega_0\, A\, \sin\omega_0 t + \omega_0\, B\, \cos\omega_0 t) - \omega_0^2\, t\,(A\, \cos\omega_0 t + B\, \sin\omega_0 t) \tag{6.31}$$

and so

$$\ddot{x}_1^{p_1} + \omega_0^2\, x_1^{p_1} = 2\,(-\,\omega_0\, A\, \sin\omega_0 t + \omega_0\, B\, \cos\omega_0 t) \tag{6.32}$$

which should be equal to the resonant part $\frac{1}{8}\,\omega_0^2\,(C_1^2 + C_2^2)\,(C_1\, \cos\omega_0 t + C_2\, \sin\omega_0 t)$, resulting in

$$A = -\,\frac{1}{16}\,\omega_0\,(C_1^2 + C_2^2)\,C_2 \qquad \text{and} \qquad B = \frac{1}{16}\,\omega_0\,(C_1^2 + C_2^2)\,C_1\,. \tag{6.33}$$

Thus, we have

$$x_1^{p_1} = \frac{1}{16}\,(C_1^2 + C_2^2)\,\omega_0 t\,(C_1\, \sin\omega_0 t - C_2\, \cos\omega_0 t)\,. \tag{6.34}$$

Similarly, let the solution for the non-resonant forcing be $x_1^{p_2} = A\,\cos 3\omega_0 t + B\,\sin 3\omega_0 t$, one may find

$$x_1^{p_2} = -\,\frac{1}{192}\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\omega_0 t - \frac{1}{192}\,(3\,C_1^2 - C_2^2)\,C_2\,\sin 3\omega_0 t\,. \tag{6.35}$$

Finally, we obtain

$$\begin{aligned} x_1(t) &= x_1^h + x_1^{p_1} + x_1^{p_2} \\ &= A_1\,\cos\omega_0 t + B_1\,\sin\omega_0 t \\ &\quad + \frac{1}{16}\,(C_1^2 + C_2^2)\,\omega_0 t\,(C_1\,\sin\omega_0 t - C_2\,\cos\omega_0 t) \\ &\quad - \frac{1}{192}\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\omega_0 t - \frac{1}{192}\,(3\,C_1^2 - C_2^2)\,C_2\,\sin 3\omega_0 t \end{aligned} \tag{6.36}$$

which becomes

$$\begin{aligned} x_1(t) &= \frac{1}{192}\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos\omega_0 t\; - \frac{1}{64}\,(7\,C_1^2 + 3\,C_2^2)\,C_2\,\sin\omega_0 t \\ &\quad + \frac{1}{16}\,(C_1^2 + C_2^2)\,\omega_0 t\,(C_1\,\sin\omega_0 t - C_2\,\cos\omega_0 t) \\ &\quad - \frac{1}{192}\,\Big[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\omega_0 t + (3\,C_1^2 - C_2^2)\,C_2\,\sin 3\omega_0 t\,\Big] \end{aligned} \tag{6.37}$$

when the initial conditions $x_1(0) = 0$ and $\dot{x}_1(0) = 0$ are used. Thus, the solution of the nonlinear DE (6.11) is approximated by

$$\begin{aligned} x(t,\,\epsilon) &= C_1\,\cos\omega_0 t + C_2\,\sin\omega_0 t \\ &\quad + \frac{1}{192}\,\epsilon\,\Big[(C_1^2 - 3\,C_2^2)\,C_1\,\cos\omega_0 t + 3\,(7\,C_1^2 + 3\,C_2^2)\,C_2\,\sin\omega_0 t\,\Big] \\ &\quad + \frac{1}{16}\,\epsilon\,(C_1^2 + C_2^2)\,\omega_0\,(C_1\,t\,\sin\omega_0 t - C_2\,t\,\cos\omega_0 t) \\ &\quad - \frac{1}{192}\,\epsilon\,\Big[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\omega_0 t + (3\,C_1^2 - C_2^2)\,C_2\,\sin 3\omega_0 t\,\Big] + O(\epsilon^2 t^2)\,. \end{aligned} \tag{6.38}$$

The terms given in the above expression including $t\,\cos t$ and $t\,\sin t$ are called *secular terms*, i.e., terms that *grow* without bound as $t \to \infty$.

For any *fixed* t, the above expression (series) (6.38) gives a good approximation as long as ϵ is small enough, in particular, $\epsilon\,t \ll 1$ so that the correction term (which is actually $O(\epsilon^2 t^2)$) is negligible. However, normally, we are interested in the behavior of the system

for *fixed* ϵ, not fixed t. Then, the above solution is only valid when $t \ll O(\frac{1}{\epsilon})$. The **Poincare expansion** theorem tells us that this series must converge when $t \in (0, T)$ and $\epsilon \in (0, \epsilon_0)$ for some T and ϵ_0. It does not guarantee *uniform convergence* for all t. For t small compared to a period 2π, the series should be quite useful, even though we expect periodic behavior for a pendulum. But because of non-periodic *secular* terms that grow in an unbounded fashion with time, the behavior of the system described by the approximate solution for a long time period will contradict to observed periodic motions. This implies that when time increases, we have to add on more terms to keep the error down. This is the characteristic of non-uniform convergence. Obviously, this is at least a nuisance, and certainly it obscures the true periodic behavior solutions. Moreover, in practice, solutions up to lower order perturbation equations can be easily obtained. Thus, it is necessary to develop some methodology which may get rid of the trouble caused by secular terms.

6.3 Lindstedt-Poincare Procedure

We have seen in applying the direct perturbation procedure that the frequency of the motion is fixed at the natural frequency, ω_0, even when $\epsilon \neq 0$. In fact, the frequency of the motion should be varied with respect to ϵ. The main idea of the Lindstedt-Poincare Procedure is to modify the frequency of the motion as a function of ϵ so that the *secular* terms may be eliminated. To achieve this, first let

$$\omega(\epsilon) = \omega_0 + \epsilon\,\omega_1 + \epsilon^2\,\omega_2 + \ldots \tag{6.39}$$

and then let

$$\tau = \omega(\epsilon)\,t \tag{6.40}$$

which is usually called *time scale*. Next assume the solution of (6.11) is in the form

$$x(\tau, \epsilon) = x_0(\tau) + \epsilon\,x_1(\tau) + \epsilon^2\,x_2(\tau) + \ldots \tag{6.41}$$

Then with the relations

$$\begin{aligned} \frac{\partial x}{\partial t} &= \frac{\partial x}{\partial \tau}\frac{\partial \tau}{\partial t} = \omega(\epsilon)\frac{\partial x}{\partial \tau} \\ \frac{\partial^2 x}{\partial t^2} &= \frac{\partial}{\partial t}\left(\frac{\partial x}{\partial t}\right) = \frac{\partial}{\partial \tau}\left(\omega(\epsilon)\frac{\partial x}{\partial \tau}\right)\frac{\partial x}{\partial \tau} = \omega^2(\epsilon)\frac{\partial^2 x}{\partial \tau^2}\,, \end{aligned} \tag{6.42}$$

the original DE (6.11) can be written as

$$\omega^2(\epsilon)\,\frac{\partial^2 x}{\partial^2 \tau} + \omega_0^2\,x - \frac{1}{6}\,\epsilon\,\omega_0^2 x^3 = 0 \tag{6.43}$$

$\Longrightarrow$

$$\begin{aligned}
&(\omega_0 + \epsilon\,\omega_1 + \epsilon^2\,\omega_2 + \ldots)^2 \left(\frac{\partial^2 x_0}{\partial \tau^2} + \epsilon\,\frac{\partial^2 x_1}{\partial \tau^2} + \epsilon^2\,\frac{\partial^2 x_2}{\partial \tau^2} + \ldots\right) \\
+ \;& \omega_0^2\,(x_0 + \epsilon\,x_1 + \epsilon^2\,x_2 + \ldots) - \frac{1}{6}\,\omega_0^2\,(\epsilon\,x_0^3 + 3\,\epsilon^2\,x_0^2 x_1 + \cdots) = 0
\end{aligned} \tag{6.44}$$

$\Longrightarrow$

$$\begin{aligned}
&\omega_0^2\,\frac{\partial^2 x_0}{\partial \tau^2} + \epsilon\,\left(\omega_0^2\,\frac{\partial^2 x_1}{\partial \tau^2} + 2\,\omega_0\,\omega_1\,\frac{\partial^2 x_0}{\partial \tau^2}\right) \\
+ \;& \epsilon^2\,\left(\omega_0^2\,\frac{\partial^2 x_2}{\partial \tau^2} + 2\,\omega_0\,\omega_1\,\frac{\partial^2 x_1}{\partial \tau^2} + (\omega_1^2 + 2\,\omega_0\,\omega_2)\,\frac{\partial^2 x_0}{\partial \tau^2}\right) + \ldots \\
+ \;& \omega_0^2\,(x_0 + \epsilon\,x_1 + \epsilon^2\,x_2 + \ldots) - \frac{1}{6}\,\omega_0^2\,(\epsilon\,x_0^3 + 3\,\epsilon^2\,x_0^2 x_1 + \ldots) = 0\,.
\end{aligned} \tag{6.45}$$

By balancing the coefficients of like powers of ϵ in the above equation, one obtains

$$\epsilon^0 \;:\quad \omega_0^2\,\frac{\partial^2 x_0}{\partial \tau^2} + \omega_0^2\,x_0 = 0 \tag{6.46}$$

$$\epsilon^1 \;:\quad \omega_0^2\,\frac{\partial^2 x_1}{\partial \tau^2} + \omega_0^2\,x_1 = -2\,\omega_0\,\omega_1\,\frac{\partial^2 x_0}{\partial \tau^2} + \frac{1}{6}\,\omega_0^2\,x_0^3 \tag{6.47}$$

$$\epsilon^2 \;:\quad \omega_0^2\,\frac{\partial^2 x_2}{\partial \tau^2} + \omega_0^2\,x_2 = -2\,\omega_0\,\omega_1\,\frac{\partial^2 x_1}{\partial \tau^2} - (\omega_1^2 + 2\,\omega_0\,\omega_2)\,\frac{\partial^2 x_0}{\partial \tau^2} + \frac{1}{2}\,\omega_0^2\,x_0^2\,x_1 \tag{6.48}$$

etc.

It can be seen by comparing the above set of linear DEs with those obtained from the direct perturbation procedure that now RHS has additional terms contributed from frequency $\omega(\epsilon)$.

The initial conditions of the above linear d.e's are determined from $x(0) = C_1$ and $\dot{x}(0) = C_2$. Since $t = 0 \Longleftrightarrow \tau = 0$, so $x(t=0) \Longleftrightarrow x(\tau = 0)$ which yields

$$x(t)|_{t=0} = C_1 = x(\tau)|_{\tau=0} = x_0(0) + \epsilon\,x_1(0) + \epsilon^2\,x_2(0) + \cdots\,. \tag{6.49}$$

By balancing the coefficients of like powers of ϵ, we have

$$x_0(0) = C_1 \qquad \text{and} \qquad x_i(0) = 0 \quad \forall\ i \geq 1\,.$$

Similarly, by balancing the first derivatives we may find

$$\begin{aligned}
&\frac{dx(t)}{dt}|_{t=0} = \omega_0\, C_2 = \omega(\epsilon)\, \frac{\partial x(\tau)}{\partial \tau}|_{\tau=0} \\
= \;& \left(\omega_0 + \epsilon\,\omega_1 + \epsilon^2\,\omega_2 + \cdots\right)\left(\frac{\partial x_0(0)}{\partial \tau} + \epsilon\,\frac{\partial x_1(0)}{\partial \tau} + \epsilon^2\,\frac{\partial x_2(0)}{\partial \tau} + \cdots\right) \\
= \;& \omega_0\,\frac{\partial x_0(0)}{\partial \tau} + \epsilon\left(\omega_0\,\frac{\partial x_1(0)}{\partial \tau} + \omega_1\,\frac{\partial x_0(0)}{\partial \tau}\right) \\
&\qquad + \epsilon^2\left(\omega_0\,\frac{\partial x_2(0)}{\partial \tau} + \omega_1\,\frac{\partial x_1(0)}{\partial \tau} + \omega_2\,\frac{\partial x_0(0)}{\partial \tau}\right) + \cdots
\end{aligned} \tag{6.50}$$

which results in, by balancing the coefficients of ϵ

$$\begin{aligned}
\frac{\partial x_0(0)}{\partial \tau} &= C_2 \\
\frac{\partial x_1(0)}{\partial \tau} &= -\left(\frac{\omega_1}{\omega_0}\right)\frac{\partial x_0(0)}{\partial \tau} = -\left(\frac{\omega_1}{\omega_0}\right) C_2 \\
\frac{\partial x_2(0)}{\partial \tau} &= -\left(\frac{\omega_1}{\omega_0}\right)\frac{\partial x_1(0)}{\partial \tau} - \left(\frac{\omega_2}{\omega_0}\right)\frac{\partial x_0(0)}{\partial \tau} = \left(\frac{\omega_1^2 - \omega_0\,\omega_2}{\omega_0^2}\right) C_2 \\
&\vdots
\end{aligned} \tag{6.51}$$

Solving the ϵ^0-order perturbation equation (6.46) yields

$$x_0(\tau) = C_1\,\cos\tau + C_2\,\sin\tau \tag{6.52}$$

where the initial conditions (6.50) have been used. With this solution, the ϵ^1-order perturbation equation (6.47) becomes

$$\begin{aligned}
\omega_0^2\,\frac{\partial^2 x_1}{\partial \tau^2} + \omega_0^2\, x_1 &= \omega_0\left[2\,\omega_1 + \frac{1}{8}\,\omega_0\,(C_1^2 + C_2^2)\right]\,(C_1\,\cos\tau + C_2\,\sin\tau) \\
&\quad + \frac{1}{24}\,\omega_0^2\left[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\tau + (3\,C_1^2 - C_2^2)\,C_2\,\sin 3\tau\,\right].
\end{aligned} \tag{6.53}$$

In order to eliminate the term $(C_1\,\cos\tau + C_2\,\sin\tau)$ which will produce secular terms in solution x_1, the coefficient of this term must be zero, which results in

$$\omega_1 = -\frac{1}{16}\,\omega_0\,(C_1^2 + C_2^2)\,, \tag{6.54}$$

and thus the solution for $x_1(\tau)$ can be found as

$$x_1(\tau) = A\,\cos\tau + B\,\sin\tau - \frac{1}{192}\left[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\tau + (3\,C_1^2 - C_2^2)\,C_2\,\sin 3\tau\,\right] \tag{6.55}$$

where A and B are arbitrary constants determined from the initial conditions. Applying the initial conditions $x_1(0) = 0$ and $\dfrac{\partial x_1(0)}{\partial \tau} = -\left(\dfrac{\omega_1}{\omega_0}\right) C_2 = \dfrac{1}{16}\,(C_1^2 + C_2^2)\,C_2$ to (6.55) and using the method of *undetermined coefficients* yield

$$A = \frac{1}{192}\,(C_1^2 - 3\,C_2^2)\,C_1 \qquad \text{and} \qquad B = \frac{1}{64}\,(7\,C_1^2 + 3\,C_2^2)\,C_2. \tag{6.56}$$

Thus,

$$\begin{aligned} x_1(\tau) &= \frac{1}{192}\,\Big[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos\tau + 3\,(7\,C_1^2 + 3\,C_2^2)\,C_2\,\sin\tau\,\Big] \\ &- \frac{1}{192}\,\Big[\,(C_1^2 - 3\,C_2^2)\,C_1\,\cos 3\tau + (3\,C_1^2 - C_2^2)\,C_2\,\sin 3\tau\,\Big]. \end{aligned} \tag{6.57}$$

The similar process can be applied to the ϵ^2-order perturbation equation to obtain

$$\begin{aligned} \omega_0^2\,\frac{\partial^2 x_2}{\partial \tau^2} + \omega_0^2\,x_2 &= \omega_0\left[2\,\omega_2 + \frac{\omega_0}{1536}\,(7\,C_1^4 + 46\,C_1^2C_2^2 + 23\,C_2^4)\right](C_1\,\cos\tau + C_2\,\sin\tau) \\ &+ \frac{1}{384}\,\omega_0^2\,\Big[(2\,C_1^4 - 15\,C_1^2C_2^2 - 9\,C_2^4)\,C_1\,\cos 3\tau \\ &\qquad + (11\,C_1^4 - C_1^2C_2^2 - 4\,C_2^4)\,C_2\,\sin 3\tau \\ &- \frac{1}{1536}\,\omega_0^2\,\Big[(C_1^4 - 10\,C_1^2\,C_2^2 + 5\,C_2^4)\,C_1\,\cos 5\tau \\ &\qquad + (5\,C_1^4 - 10\,C_1^2\,C_2^2 + C_2^4)\,C_2\,\sin 5\tau\Big] \end{aligned} \tag{6.58}$$

which in turn yields

$$\omega_2 = -\,\frac{\omega_0}{3072}\,(7\,C_1^4 + 46\,C_1^2C_2^2 + 23\,C_2^4) \tag{6.59}$$

by eliminating the secular terms which may appear in x_2; and thus

$$\begin{aligned} x_2(\tau) &= \frac{1}{36864}\,\Big[(23C_1^4 - 170C_1^2C_2^2 - 113C_2^4)C_1\,\cos\tau + (599C_1^4 + 854C_1^2C_2^2 + 271C_2^4)C_2\,\sin\tau\Big] \\ &- \frac{1}{3072}\,\Big[(2\,C_1^4 - 15\,C_1^2C_2^2 - 9\,C_2^4)\,C_1\,\cos 3\tau + (11\,C_1^4 - C_1^2C_2^2 - 4\,C_2^4)\,C_2\,\sin 3\tau\Big] \\ &+ \frac{1}{36864}\,\Big[(C_1^4 - 10\,C_1^2\,C_2^2 + 5\,C_2^4)\,C_1\,\cos 5\tau + (5\,C_1^4 - 10\,C_1^2\,C_2^2 + C_2^4)\,C_2\,\sin 5\tau\Big] \end{aligned} \tag{6.60}$$

where the initial conditions $x_2(0) = 0$ and $\dfrac{\partial x_2(0)}{\partial \tau} = \left(\dfrac{\omega_1^2 - \omega_0\,\omega_2}{\omega_0^2}\right) C_2 = \dfrac{1}{3072}\,(19\,C_1^4 + 70\,C_1^2\,C_2^2 + 25\,C_2^4)\,C_2$ have been used.

Hence, the approximate solution of $x(\tau,\epsilon)$ up to second order is, therefore, given by

$$\begin{aligned} x(\tau,\epsilon) &= C_1\cos\tau + C_2\sin\tau \\ &\quad + \frac{1}{192}\epsilon\Big[(C_1^2-3\,C_2^2)C_1\cos\tau + 3\,(7\,C_1^2+3\,C_2^2)C_2\sin\tau \\ &\qquad\qquad - (C_1^2-3\,C_2^2)C_1\cos 3\tau - (3\,C_1^2-C_2^2)C_2\sin 3\tau\Big] \\ &\quad + \frac{1}{36864}\epsilon^2\Big[(23\,C_1^4-170\,C_1^2C_2^2-113\,C_2^4)\,C_1\cos\tau \\ &\qquad\qquad + (599\,C_1^4+854\,C_1^2C_2^2+271\,C_2^4)\,C_2\sin\tau \\ &\qquad\qquad - 12\,(2\,C_1^4-15\,C_1^2C_2^2-9\,C_2^2)\,C_1\cos 3\tau \\ &\qquad\qquad - 12\,(11\,C_1^4-C_1^2C_2^2-4\,C_2^4)\,C_2\sin 3\tau \\ &\qquad\qquad + (C_1^4-10\,C_1^2\,C_2^2+5\,C_2^4)\,C_1\cos 5\tau \\ &\qquad\qquad + (5\,C_1^4-10\,C_1^2\,C_2^2+C_2^4)\,C_2\sin 5\tau\Big] + O(\epsilon^3)\,. \end{aligned} \tag{6.61}$$

where

$$\omega(\epsilon) = \omega_0\left[1-\frac{\epsilon}{16}\,(C_1^2+C_2^2) - \frac{\epsilon^2}{3072}\,(7\,C_1^4+46\,C_1^2C_2^2+23\,C_2^4)\right] + O(\epsilon^3)\,. \tag{6.62}$$

Now, this series is a truly periodic solution for any large t.

6.4 Time Averaging (K-B Method)

For weakly nonlinear DEs, given in the form

$$\ddot{x} + \omega_0^2\,x = \epsilon\,f(x,\,\dot{x},\,\epsilon) \tag{6.63}$$

where function f contains nonlinear terms, the **time averaging** method may be the simplest perturbation method in finding *first-order approximate solutions.* The idea of this method is again based on the perturbation of the solutions of the corresponding linear DE For $\epsilon = 0$, the solution of the linear DE is described by

$$x(t) = a\,\cos(\omega_0 t + \phi) \tag{6.64}$$

where a and ϕ are constants determined from the initial conditions. Note that an alternative solution form different from the previous sections is used here. For $\epsilon \neq 0$, it is assumed that

the solution is still in the same form, but a and ϕ are now functions of t, i.e.,

$$x(t) = a(t)\,\cos\left[\omega_0 t + \phi(t)\right] \tag{6.65}$$

which is called *parameter variation*. Then the derivative of $x(t)$ is

$$\begin{aligned} \dot{x}(t) &= \dot{a}(t)\,\cos\left[\,\omega_0 t + \phi(t)\right] - a(t)\,\sin\left[\,\omega_0 t + \phi(t)\right]\,\left[\omega_0 + \dot{\phi}(t)\right] \\ &= -\,\omega_0\,a\,\sin(\omega_0 t + \phi) + \dot{a}\,\cos(\omega_0 t + \phi) - a\,\dot{\phi}\,\sin(\omega_0 t + \phi)\,. \end{aligned} \tag{6.66}$$

Since now there are two extra variables a and ϕ to be determined, it is possible to set

$$\dot{a}\,\cos(\omega_0 t + \phi) - a\,\dot{\phi}\,\sin(\omega_0 t + \phi) = 0 \tag{6.67}$$

so that

$$\dot{x} = -\,\omega_0\,a\,\sin(\omega_0 t + \phi)\,. \tag{6.68}$$

Then the second derivative of $\dot{x}$ can be written as

$$\begin{aligned} \ddot{x} &= -\,\omega_0\,\left[\,\dot{a}\,\sin(\omega_0 t + \phi) + a\,\cos(\omega_0 t + \phi)\,(\omega_0 + \dot{\phi})\,\right] \\ &= -\,\omega_0^2\,a\,\cos(\omega_0 t + \phi) - \omega_0\,\dot{a}\,\sin(\omega_0 t + \phi) - \omega_0\,a\,\dot{\phi}\,\cos(\omega_0 t + \phi) \end{aligned} \tag{6.69}$$

which, in turn, yields

$$\begin{aligned} \dot{a}\,\sin(\omega_0 t + \phi) + a\,\dot{\phi}\,\cos(\omega_0 t + \phi) &= -\,\frac{1}{\omega_0}\,\left[\,\ddot{x} + \omega_0^2\,a\,\cos(\omega_0 t + \phi t)\,\right] \\ &= -\,\frac{1}{\omega_0}\,(\ddot{x} + \omega_0^2\,x) \\ &= -\,\frac{1}{\omega_0}\,\epsilon\,f(x,\,\dot{x},\,\epsilon) \end{aligned} \tag{6.70}$$

which can solved together with the equation (6.67) for $\dot{a}$ and $\dot{\phi}$ to obtain

$$\begin{aligned} \dot{a} &= -\,\frac{\epsilon}{\omega_0}\,f(x,\,\dot{x},\,\epsilon)\,\sin(\omega_0 t + \phi) \\ a\,\dot{\phi} &= -\,\frac{\epsilon}{\omega_0}\,f(x,\,\dot{x},\,\epsilon)\,\cos(\omega_0 t + \phi) \end{aligned} \tag{6.71}$$

by a simple algebraic manipulation. Since $\dot{a}$ and $\dot{\phi}$ are produced from small ϵ, they can be considered as small (usually called *slow time variables*), and so they may be treated as

constants in one period. Therefore integrating the above equation in one period gives

$$\begin{aligned} \dot{a} &\approx -(\frac{\epsilon}{\omega_0})(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} f(x,\,\dot{x},\,\epsilon)\,\sin(\omega_0 t+\phi)\,dt \\ a\,\dot{\phi} &\approx -(\frac{\epsilon}{\omega_0})(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} f(x,\,\dot{x},\,\epsilon)\,\cos(\omega_0 t+\phi)\,dt \end{aligned} \tag{6.72}$$

These two equations are called **Averaged Equations**.

To find the explicit expressions of the integrals, one may expand function f in *Fourier Series*:

$$f(x,\,\dot{x},\,\epsilon) = \sum_{k=0}^{\infty} [\,p_k\,\cos k(\omega_0 t+\phi) + q_k\,\sin k(\omega_0 t+\phi)\,] \tag{6.73}$$

where the Fourier coefficients p_k and q_k are given by

$$\begin{aligned} p_0 &= \frac{\omega_0}{2\,\pi} \int_0^{\frac{2\,\pi}{\omega_0}} f(x,\,\dot{x},\,\epsilon)\,dt \\ p_k &= \frac{\omega_0}{\pi} \int_0^{\frac{2\,\pi}{\omega_0}} f(x,\,\dot{x},\,\epsilon)\,\cos k(\omega_0 t+\phi)\,dt \qquad \text{for } k\geq 1 \\ q_k &= \frac{\omega_0}{\pi} \int_0^{\frac{2\,\pi}{\omega_0}} f(x,\,\dot{x},\,\epsilon)\,\sin k(\omega_0 t+\phi)\,dt \end{aligned} \tag{6.74}$$

Therefore,

$$\begin{aligned} \dot{a} &= -(\frac{\epsilon}{\omega_0})\,(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} \sum_{k=0}^{\infty} [\,p_k\,\cos k(\omega_0 t+\phi) + q_k\,\sin k(\omega_0 t+\phi)\,]\,\sin(\omega_0 t+\phi)\,dt \\ &= -(\frac{\epsilon}{\omega_0^2})\,(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} q_1\,\sin^2(\omega_0 t+\phi)\,d(\omega_0 t+\phi) \\ &= -(\frac{\epsilon}{2\,\omega_0^2})\,q_1\,(a,\,\phi) \end{aligned} \tag{6.75}$$

and

$$\begin{aligned} a\,\dot{\phi} &= -(\frac{\epsilon}{\omega_0})\,(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} \sum_{k=0}^{\infty} [\,p_k\,\cos k(\omega_0 t+\phi) + q_k\,\sin k(\omega_0 t+\phi)\,]\,\sin k(\omega_0 t+\phi)\,dt \\ &= -(\frac{\epsilon}{\omega_0^2})\,(\frac{\omega_0}{2\,\pi}) \int_0^{\frac{2\,\pi}{\omega_0}} p_1\,\cos^2(\omega_0 t+\phi)\,d(\omega_0 t+\phi) \\ &= -(\frac{\epsilon}{2\,\omega_0^2})\,p_1\,(a,\,\phi) \end{aligned} \tag{6.76}$$

where p_1 and q_1 are functions of a and ϕ. Setting $\dot{a}=\dot{\phi}=0$ and solving the resulting equations produce *steady-state* (asymptotic) solutions $\bar{a}$, $\bar{\phi}$, which is periodic motion. The first-order approximation of the DE (6.63) is, therefore, described by

$$x(t) = \bar{a}\,\cos(\omega_0 t+\bar{\phi})\,. \tag{6.77}$$

Furthermore, based on the averaged equation, one may carry out a stability analysis for the steady-state solutions, which is demonstrated in the following example.

EXAMPLE 6.1. Consider the *van der Pol* equation:

$$\ddot{x} + x + \epsilon\,(x^2 - 1)\,\dot{x} = 0\,. \tag{6.78}$$

In this case, $\omega_0 = 1$ and $f(x,\,\dot{x},\,\epsilon) = -\,(x^2 - 1)\,\dot{x}$. Applying the above formulas (6.65), (6.68) and (6.72) (and remember treating a and ϕ as constants) produces

$$\dot{a} \;=\; \frac{1}{2}\epsilon\; a\,(1 - \frac{1}{4}\,a^2) \tag{6.79}$$

$$a\,\dot{\phi} \;=\; 0 \tag{6.80}$$

Setting $\dot{a} = 0$ gives $\bar{a} = 2$ and (6.80) (which satisfies $\dot{\phi} = 0$.) yields $\bar{\phi} = \phi_0$, and thus the first-order steady-state solution is

$$x(t) = 2\,\cos(t + \phi_0) \tag{6.81}$$

where ϕ_0 is a constant phase.

To perform the stability analysis, one needs the averaged equation $\dot{a}$ only since a is decoupled from ϕ in this case. The stability analysis is similar to analyzing the stability of fixed points of DEs. The derivative of $\dot{a}$ with respect to a is equal to

$$J = \frac{\partial \dot{a}}{\partial a} = \frac{1}{2}\,\epsilon\,\left(1 - \frac{3}{4}\,a^2\right) \qquad \text{and so} \qquad J|_{a\,=\,\bar{a}\,=\,2} = -\,\epsilon \quad (\epsilon > 0)\,. \tag{6.82}$$

Therefore, the steady state solution is stable. In other words, the van der Pol equation exhibits stable limit cycles (nonlinear oscillations).

6.5 Multiple Time Scales

The *Multiple Time Scales* or simply *Multiple scales* is frequently used in analyzing nonlinear oscillations. The underlying idea of the method is to consider the expansion representing the response to be a function of multiple independent variables, or scales, instead of a single time variable.

One begins by introducing new independent variables according to

$$T_n = \epsilon^n t \qquad \text{for} \qquad n = 0,\, 1,\, 2,\, \ldots \tag{6.83}$$

It follows that the derivatives with respect to t become expansions in terms of the partial derivatives with respect to T_n according to

$$\begin{aligned}
\frac{d}{dt} &= \frac{\partial}{\partial T_0}\frac{dT_0}{dt} + \frac{\partial}{\partial T_1}\frac{dT_1}{dt} + \frac{\partial}{\partial T_2}\frac{dT_2}{dt} + \ldots = D_0 + \epsilon\, D_1 + \epsilon^2\, D_2 + \ldots \\
\frac{d^2}{dt^2} = \frac{d}{dt}\left(\frac{d}{dt}\right) &= \frac{d}{dt}\left(\frac{\partial}{\partial T_0} + \epsilon\,\frac{\partial}{\partial T_1} + \epsilon^2\,\frac{\partial}{\partial T_2} + \cdots\right) \\
&= \frac{\partial}{\partial T_0}\left(\frac{\partial}{\partial T_0}\right)\frac{dT_0}{dt} + \frac{\partial}{\partial T_1}\left(\frac{\partial}{\partial T_0}\right)\frac{dT_1}{dt} + \frac{\partial}{\partial T_2}\left(\frac{\partial}{\partial T_0}\right)\frac{dT_2}{dt} \\
&\quad + \epsilon\left[\frac{\partial}{\partial T_0}\left(\frac{\partial}{\partial T_1}\right)\frac{dT_0}{dt} + \frac{\partial}{\partial T_1}\left(\frac{\partial}{\partial T_1}\right)\frac{dT_1}{dt}\right] + \epsilon^2\,\frac{\partial}{\partial T_0}\left(\frac{\partial}{\partial T_2}\right)\frac{dT_0}{dt} + \cdots \\
&= D_0^2 + 2\,\epsilon\, D_1\, D_0 + \epsilon^2\,(D_1^2 + 2\, D_2\, D_0) + \ldots
\end{aligned} \tag{6.84}$$

where $D_i = \dfrac{\partial}{\partial T_i},\ i = 1,\, 2,\, 3,\, \cdots$ One then assumes that the solution of a nonlinear DE can be represented by an expansion in the form

$$x(t;\epsilon) = x_0(T_0,\, T_1,\, T_2,\, \ldots) + \epsilon\, x_1(T_0,\, T_1,\, T_2,\, \ldots) + \epsilon^2\, x_2(T_0,\, T_1,\, T_2,\, \ldots) + \ldots \tag{6.85}$$

We note that the number of independent time scales needed depends on the order to which the expansion is carried out. If the expansion is carried out to $O(\epsilon^2)$, then T_0, T_1 and T_2 are needed. In this section, we carry out the expansion to $O(\epsilon^2)$, and hence we need T_0, T_1 and T_2. We shall use the equation of the real pendulum to illustrate the application of the multiple time scales. We rewrite the DE (6.11) here for convenience

$$\ddot{x} + \omega_0^2\, x - \frac{1}{6}\,\epsilon\,\omega_0^2\, x^3 = 0\,.$$

Substituting

$$\begin{aligned}
\ddot{x} &= \left[D_0^2 + 2\,\epsilon\, D_0\, D_1 + \epsilon^2\,(D_1^2 + 2\, D_0\, D_1) + \ldots\right](x_0 + \epsilon\, x_1 + \epsilon^2\, x_2 + \ldots) \\
&= D_0^2\, x_0 + \epsilon\,(D_0^2\, x_1 + 2\, D_0\, D_1\, x_0) + \epsilon^2\,\left[D_0^2\, x_2 + 2\, D_0\, D_1\, x_1 (D_1^2 + 2\, D_0\, D_2)\, x_0\right] + \ldots
\end{aligned} \tag{6.86}$$

and x into (6.11) and then balancing the coefficients of like powers of ϵ yields

$$\epsilon^0 \; : \quad D_0^2\, x_0 + \omega_0^2\, x_0 = 0 \tag{6.87}$$

$$\epsilon^1 \; : \quad D_0^2\, x_1 + \omega_0^2\, x_1 = -2\, D_1\, D_0\, x_0 + \frac{1}{6}\,\omega_0^2\, x_0^3 \tag{6.88}$$

$$\epsilon^2 \; : \quad D_0^2\, x_2 + \omega_0^2\, x_2 = -2\, D_1\, D_0\, x_1 - (D_1^2 + 2\, D_2\, D_0)\, x_0 + \frac{1}{2}\,\omega_0^2\, x_0^2\, x_1 \tag{6.89}$$

etc.

The solution for the ϵ^0-order equation can be written as

$$x_0 = a(T_1,\, T_2)\, \cos\left[\, \omega_0 T_0 + \phi(T_1,\, T_2)\,\right] \equiv a\, \cos(\omega_0 T_0 + \phi) \tag{6.90}$$

where a and ϕ represent the amplitude and phase of the motion, respectively. Then by using this solution, we have

$$\begin{aligned} D_0\, D_1\, x_0 &= D_1\, (D_0\, x_0) \\ &= D_1\, \left[-\,\omega_0\, a\, \sin(\omega_0 T_0 + \phi)\right] \\ &= -\,\omega_0\, D_1 a\, \sin(\omega_0 T_0 + \phi) - \omega_0\, a\, D_1\phi\, \cos(\omega_0 T_0 + \phi)\,. \end{aligned} \tag{6.91}$$

Thus the ϵ^1-order equation becomes

$$\begin{aligned} & D_0^2\, x_1 + \omega_0^2\, x_1 \\ = \;& 2\,\omega_0\, D_1 a\, \sin(\omega_0 T_0 + \phi) + 2\,\omega_0\, a\, D_1\phi\, \cos(\omega_0 T_0 + \phi) + \frac{1}{6}\,\omega_0^2\, a^3\, \cos^3(\omega_0 T_0 + \phi) \\ = \;& 2\,\omega_0\, D_1 a\, \sin(\omega_0 T_0 + \phi) + 2\,\omega_0\, a\, D_1\phi\, \cos(\omega_0 T_0 + \phi) \\ & + \left(\frac{1}{6}\,\omega_0^2\, a^3\right)\, \frac{1}{4}\, \left[\, 3\, \cos(\omega_0 T_0 + \phi) + \cos 3(\omega_0 T_0 + \phi)\,\right] \\ = \;& 2\,\omega_0\, D_1 a\, \sin(\omega_0 T_0 + \phi) + \left[2\,\omega_0\, a\, D_1\phi + \frac{1}{8}\,\omega_0^2\, a^3\right]\, \cos(\omega_0 T_0 + \phi) \\ & + \frac{1}{24}\,\omega_0^2\, a^3\, \cos 3(\omega_0 T_0 + \phi)\,. \end{aligned} \tag{6.92}$$

Eliminating the secular terms in the solution of x_1 requires

$$\begin{aligned} D_1\, a &= 0 \\ D_1\, \phi &= -\,\frac{1}{16}\,\omega_0\, a^2 \end{aligned} \tag{6.93}$$

and thus

$$D_0^2\, x_1 + x_1 = \frac{1}{24}\,\omega_0^2\, a^3\, \cos 3(\omega_0 T_0 + \phi) \tag{6.94}$$

which has the *particular* solution (NOTE: For the MTS method, we only need to find the particular solution, no homogeneous solution)

$$x_1 = -\,\frac{1}{192}\, a^3\, \cos 3(\omega_0 T_0 + \phi)\,. \tag{6.95}$$

Next, to find the solution of the ϵ^2-order equation, one needs to first calculate the following terms needed in the equation:

$$\begin{aligned} D_1\, D_0\, x_1 &= D_1\,(D_0\, x_1) \\ &= \frac{1}{64}\,\omega_0\; D_1\left[\, a^3\, \sin 3(\omega_0 T_0 + \phi)\,\right] \\ &= \frac{1}{64}\,\omega_0\; D_1\left[\, 3\, a^2\, D_1 a\, \sin 3(\omega_0 T_0 + \phi) + 3\, a^3\, D_1\phi\, \cos 3(\omega_0 T_0 + \phi)\,\right] \\ &= -\,\frac{3}{1024}\,\omega_0^2\, a^5\, \cos 3(\omega_0 T_0 + \phi) \end{aligned} \tag{6.96}$$

where $D_1\, a = 0$ and $D_1\,\phi = -\,\dfrac{1}{16}\,\omega_0\, a^2$ have been used.

$$\begin{aligned} D_1^2\, x_0 &= D_1\,(D_1\, x_0) \\ &= D_1\,[\, D_1 a\, \cos(\omega_0 T_0 + \phi) - a\, D_1\phi\, \sin(\omega_0 T_0 + \phi)\,] \\ &= \frac{1}{16}\,\omega_0\; D_1\left[\, a^3\, \sin(\omega_0 T_0 + \phi)\,\right] \\ &= \frac{1}{16}\,\omega_0\; D_1\left[\, 3\, a^2\, D_1 a \sin(\omega_0 T_0 + \phi) + a^3\, D_1\phi\, \cos(\omega_0 T_0 + \phi)\,\right] \\ &= -\,\frac{1}{256}\,\omega_0^2\, a^5\, \cos(\omega_0 T_0 + \phi) \end{aligned} \tag{6.97}$$

$$\begin{aligned} D_2\, D_0\, x_0 &= D_2\,(D_0\, x_0) \\ &= D_2\,[\, -\,\omega_0\, a\, \sin(\omega_0 T_0 + \phi)\,] \\ &= -\,\omega_0\, D_2 a\, \sin(\omega_0 T_0 + \phi) - \omega_0\, a\, D_2\phi\, \cos(\omega_0 T_0 + \phi) \end{aligned} \tag{6.98}$$

and

$$\frac{1}{2}\,\omega_0^2\, x_0^2\, x_1 = \frac{1}{2}\,\omega_0^2\, a^2\, \cos^2(\omega_0 T_0 + \phi)\; (-\,\frac{1}{192})\, a^3\, a^3\, \cos 3(\omega_0 T_0 + \phi)$$

$$
\begin{aligned}
&= -\frac{1}{768}\,\omega_0^2\,a^5\,\left[\,1+\cos 2(\omega_0 T_0+\phi)\,\right]\,\cos 3(\omega_0 T_0+\phi) \\
&= -\frac{1}{768}\,\omega_0^2\,a^5\,\left[\,\frac{1}{2}\,\cos(\omega_0 T_0+\phi)+\cos 3(\omega_0 T_0+\phi)+\frac{1}{2}\,\cos 5(\omega_0 T_0+\phi)\,\right] \\
&= -\frac{1}{1536}\,\omega_0^2\,a^5\,\cos(\omega_0 T_0+\phi) \\
&\quad -\frac{1}{768}\,\omega_0^2\,a^5\,\cos 3(\omega_0 T_0+\phi) \\
&\quad -\frac{1}{1536}\,\omega_0^2\,a^5\,\cos 5(\omega_0 T_0+\phi)\,. \qquad (6.99)
\end{aligned}
$$

Then substituting the above expressions into the ϵ^2-order equation yields

$$
\begin{aligned}
D_0^2\,x_2+\omega_0^2\,x_2 &= 2\,\omega_0\,D_2 a\,\sin(\omega_0 T_0+\phi) \\
&\quad +\left[\,2\,\omega_0\,a\,D_2\phi+\frac{5}{1536}\,\omega_0^2\,a^5\,\right]\,\cos(\omega_0 T_0+\phi) \\
&\quad -\frac{7}{1536}\,\omega_0^2\,a^5\,\cos 3(\omega_0 T_0+\phi)-\frac{1}{1536}\,\omega_0^2\,a^5\,\cos 5(\omega_0 T_0+\phi)\,. \qquad (6.100)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
D_2\,a &= 0 \\
D_2\,\phi &= -\frac{5}{3072}\,\omega_0\,a^4 \qquad (6.101)
\end{aligned}
$$

and

$$
x_2 = -\frac{7}{12288}\,a^5\,\cos 3(\omega_0 T_0+\phi)+\frac{1}{36864}\,a^5\,\cos 5(\omega_0 T_0+\phi)\,. \qquad (6.102)
$$

Finally, the approximate solution is given by

$$
\begin{aligned}
x(t,\epsilon) &= a\,\cos(\omega_0 t+\phi) \\
&\quad -\frac{1}{192}\,\epsilon\,a^3\,\cos 3(\omega_0 t+\phi) \\
&\quad -\frac{1}{36864}\,\epsilon^2\,a^5\,\left[\,21\,\cos 3(\omega_0 t+\phi)-\,\cos 5(\omega_0 t+\phi)\,\right]\,. \qquad (6.103)
\end{aligned}
$$

The amplitude a and the phase ϕ of the periodic motions may be determined from the following *rate equations* up to $O(\epsilon^2)$ order:

$$
\begin{aligned}
\frac{da}{dt} &= \frac{\partial a}{\partial T_1}\,\frac{\partial T_1}{\partial t}+\frac{\partial a}{\partial T_2}\,\frac{\partial T_2}{\partial t}+O(\epsilon^3) \\
&= \epsilon\,D_1 a+\epsilon^2\,D_2 a+O(\epsilon^3) \\
&= 0+O(\epsilon^3) \qquad (6.104)
\end{aligned}
$$

and

$$\begin{aligned}\frac{d\phi}{dt} &= \frac{\partial\phi}{\partial T_1}\frac{\partial T_1}{\partial t} + \frac{\partial\phi}{\partial T_2}\frac{\partial T_2}{\partial t} + O(\epsilon^3) \\ &= \epsilon\, D_1\phi + \epsilon^2\, D_2\phi + O(\epsilon^3) \\ &= -\frac{1}{16}\,\epsilon\,\omega_0\,a^2 - \frac{5}{3072}\,\epsilon^2\,\omega_0^2\,a^4 + O(\epsilon^3) \end{aligned} \tag{6.105}$$

and thus,

$$\begin{aligned} a &= a_0 \\ \phi &= \phi_0 - \left(\frac{1}{16}\,\epsilon\,\omega_0\,a^2 + \frac{5}{3072}\,\epsilon^2\,\omega_0^2\,a^4\right) t + O(\epsilon^3) \end{aligned} \tag{6.106}$$

up to $O(\epsilon^2)$ order, where a_0 and ϕ_0 are constants, representing the initial amplitude and phase of periodic solutions. ϕ_0 can be chosen zero because the system is autonomous. Once the a_0 and ϕ_0 are chosen, the periodic solution is uniquely determined. So, $\omega_0\,t + \phi$ can be written as

$$\omega_0\,t + \phi = \omega_0\,\left(1 - \frac{1}{16}\,\epsilon\,a^2 - \frac{5}{3072}\,\epsilon^2\,a^4\right) t + \phi_0 + O(\epsilon^3)\,. \tag{6.107}$$

It should be noted that the periodic solutions derived in this section are given in terms of the initial conditions a_0 and ϕ_0, while the solutions of the same Duffing equation found by using the Lindstedt-Poincaré procedure are given in terms of the initial conditions C_1 and C_2. It can be shown that these two solutions are equivalent because there exists a one to one mapping between $(C_1,\, C_2)$ and $(a_0,\, \phi_0)$. That is, given C_1 and C_2, a unique pair $(a_0,\, \phi_0)$ can be determined such that the two solutions are identical up to the same order.

6.6 Comparison of Perturbation Methods

We have learned four perturbation methods: (i) direct (regular) perturbation procedure, (ii) Lindstedt-Poincaré procedure, (iii) time averaging, and (iv) multiple time scales. All these methodologies are employed to find approximate periodic and quasiperiodic solutions. The direct perturbation procedure usually produces solutions which may involve secular terms, which make the solution a good approximation for very small time t only, and thus it is not practically useful. The Lindstedt-Poincaré procedure uses a time scale, given in

terms of perturbation parameter ϵ, to eliminate secular terms by choosing an appropriate frequency function and thus, the solution is truly periodic for any time $t \in [0, \infty)$. However, this procedure is only applicable in finding steady-state (periodic) solutions. Moreover, this method cannot be used for stability analysis of the steady-state solutions. The time averaging and multiple time scales, on the other hand, can be used to find approximate transient as well as steady-state solutions. Moreover, they can be used to determine the stability of steady-state solutions. This has been shown in studying van der Pol equation by using time averaging. It was seen that all solutions (trajectories), no matter starting from what initial conditions, eventually converge to the periodic solution having the amplitude of 2. In other words, the initial conditions do not affect the final state of the periodic solutions. The pendulum example, however, considered by using the Lindstedt-Poincaré and multiple time scales, has shown that every solution is periodic for any given initial conditions. The amplitude of a solution is a constant, uniquely determined from the initial conditions, while the frequency is modified as a function of the perturbation parameter ϵ. In fact, every solution represents a nonlinear center.

Now, let's use the perturbation methods to reconsider the van der Pol equation (6.78) and make a comparison. Since this system has been analyzed by using the time averaging in Section 6.4, for the simplicity and consistency with the results obtained using time averaging, we will only look for approximate solutions up to first $(O(\epsilon))$ order. The van der Pol equation is given by

$$\ddot{x} + x + \epsilon\,(x^2 - 1)\,\dot{x} = 0 \tag{6.108}$$

with initial conditions $x(0) = C_1$ and $\dot{x}(0) = C_2$.

6.6.1 Regular perturbation procedure. Since this method produces solutions including secular terms which render the solutions unbounded when $t \to \infty$, we are not interested in and omit the analysis.

6.6.2 Lindstedt-Poincare procedure. Let

$$\tau = \omega(\epsilon)\,t = (1 + \epsilon\,\omega_1 + \ldots)\,t \tag{6.109}$$

and

$$x(\tau,\epsilon) = x_0(\tau) + \epsilon\, x_1(\tau) + \ldots \tag{6.110}$$

Then with the aid of the relations $\dfrac{\partial x}{\partial t} = \omega(\epsilon)\,\dfrac{\partial x}{\partial \tau}$ and $\dfrac{\partial^2 x}{\partial t^2} = \omega^2(\epsilon)\,\dfrac{\partial^2 x}{\partial \tau^2}$, the DE (6.108) can be written as

$$\begin{aligned}
&(1+\epsilon\,\omega_1+\ldots)^2\left(\frac{\partial^2 x_0}{\partial\tau^2}+\epsilon\,\frac{\partial^2 x_1}{\partial\tau^2}+\ldots\right)+x_0+\epsilon\,x_1+\ldots \\
+\quad &\epsilon\,(1+\epsilon\,\omega_1+\ldots)\left(\frac{\partial x_0}{\partial\tau}+\epsilon\,\frac{\partial x_1}{\partial\tau}+\ldots\right)\Big[(x_0+\epsilon\,x_1+\ldots)^2-1\Big]=0
\end{aligned} \tag{6.111}$$

$\Longrightarrow$

$$\frac{\partial^2 x_0}{\partial\tau^2}+\epsilon\left(\frac{\partial^2 x_1}{\partial\tau^2}+2\,\omega_1\,\frac{\partial^2 x_0}{\partial\tau^2}\right)+x_0+\epsilon\,x_1+\ldots+\epsilon\,(x_0^2-1)\,\frac{\partial x_0}{\partial\tau}+\ldots=0 \tag{6.112}$$

$\Longrightarrow$

$$\epsilon^0\ :\quad \frac{\partial^2 x_0}{\partial\tau^2}+x_0=0 \tag{6.113}$$

$$\epsilon^1\ :\quad \frac{\partial^2 x_1}{\partial\tau^2}+x_1=-\,2\,\omega_1\,\frac{\partial^2 x_0}{\partial\tau^2}-(x_0^2-1)\,\frac{\partial x_0}{\partial\tau} \tag{6.114}$$

etc.

The initial conditions for x_0 and x_1 can be determined from $x(0)$ and $\dot{x}(0)$ as

$$\begin{aligned}
x_0(0) &= C_1\,, \qquad \frac{\partial x_0(0)}{\partial\tau} = C_2 \\
x_1(0) &= 0\,, \qquad \frac{\partial x_1(0)}{\partial\tau} = -\,\omega_1\,C_2\,.
\end{aligned} \tag{6.115}$$

Solving the ϵ^0-order perturbation equation (6.113) with the initial conditions yields

$$x_0(\tau) = C_1\,\cos\tau + C_2\,\sin\tau\,. \tag{6.116}$$

With this solution, the ϵ^1-order perturbation equation (6.114) now becomes

$$\begin{aligned}
\frac{\partial^2 x_1}{\partial\tau^2}+x_1 &= -\,2\,\omega_1\,(-\,C_1\,\cos\tau - C_2\,\sin\tau) \\
&\quad -\,(C_1^2\,\cos^2\tau + C_2^2\,\sin^2\tau + 2\,C_1\,C_2\,\sin\tau\,\cos\tau - 1)\,(-\,C_1\,\sin\tau + C_2\,\cos\tau)
\end{aligned} \tag{6.117}$$

$\Longrightarrow$

$$\begin{aligned}\frac{\partial^2 x_1}{\partial \tau^2} + x_1 &= 2\,\omega_1\,(C_1\,\cos\tau + C_2\,\sin\tau) \\ &\quad + \cos\tau\left(C_2 - \frac{1}{4}\,C_1^2\,C_2 - \frac{1}{4}\,C_2^3\right) - \sin\tau\left(C_1 - \frac{1}{4}\,C_1^3 - \frac{1}{4}\,C_1\,C_2^2\right) \\ &\quad - \cos 3\tau\left(\frac{3}{4}\,C_1^2\,C_2 - \frac{1}{4}\,C_2^3\right) + \sin 3\tau\left(\frac{1}{4}\,C_1^3 - \frac{3}{4}\,C_1\,C_2^2\right) \qquad (6.118)\end{aligned}$$

$\Longrightarrow$

$$\begin{aligned}\frac{\partial^2 x_1}{\partial \tau^2} + x_1 &= \cos\tau\left\{2\,\omega_1\,C_1 + C_2\,\left[1 - \frac{1}{4}\,(C_1^2 + C_2)\right]\right\} \\ &\quad + \sin\tau\left\{2\,\omega_1\,C_2 - C_1\,\left[1 - \frac{1}{4}\,(C_1^2 + C_2)\right]\right\} \\ &\quad - \frac{1}{4}\,\left(3\,C_1^2 - C_2^2\right)\,C_2\,\cos 3\tau + \frac{1}{4}\,\left(C_1^2 - 3\,C_2^2\right)\,C_1\,\sin 3\tau\,. \qquad (6.119)\end{aligned}$$

The first two terms on the right hand side of the above equation will produce secular terms in solution x_1. In order to eliminate the secular terms, it is required to set these two terms zero. Thus,

$$\begin{aligned}2\,\omega_1\,C_1 + C_2\,\left[1 - \frac{1}{4}\,(C_1^2 + C_2)\right] &= 0 \\ 2\,\omega_1\,C_2 - C_1\,\left[1 - \frac{1}{4}\,(C_1^2 + C_2)\right] &= 0\,. \qquad (6.120)\end{aligned}$$

Multiplying the first equation with C_2 and the second equation with $-C_1$, and add the resulting equations together yield

$$(C_1^2 + C_2^2)\left[1 - \frac{1}{4}(C_1^2 + C_2^2)\right] = 0 \qquad (6.121)$$

which, in turn, produces two solutions

$$C_1^2 + C_2^2 = 0 \qquad \text{and} \qquad C_1^2 + C_2^2 = 4\,. \qquad (6.122)$$

The first solution is, in fact, the trivial solution $x(t) \equiv 0$. The second solution represents a periodic solution having amplitude of 2. It should be noted that unlike the analysis of the pendulum example where the initial conditions C_1 and C_2 can be chosen arbitrarily, here C_1 and C_2 must be chosen such that the condition $C_1^2 + C_2^2 = 4$ is satisfied. This condition happens to be the amplitude of the steady-state solutions. Thus, a solution found by using

Lindstedt-Poincaré procedure represents steady-state only. That is, this method cannot be used to find transient solutions. For example, if $C_1 = 1$ and $C_2 = 0$, then one cannot find a solution for this problem by using the Lindstedt-Poincaré procedure. Having chosen C_1 and C_2 such that $C_1^2 + C_2^2 = 4$, ω_1 must be zero. Therefore, from the ϵ^1-order equation (6.115), we obtain

$$C_1^2 + C_2^2 = 4\,, \qquad \omega_1 = 0 \tag{6.123}$$

and

$$\frac{\partial^2 x_1}{\partial \tau^2} + x_1 = -\,\frac{1}{4}\left(3\,C_1^2 - C_2^2\right)\,C_2\,\cos 3\tau + \frac{1}{4}\left(C_1^2 - 3\,C_2^2\right)\,C_1\,\sin 3\tau \tag{6.124}$$

which gives the solution

$$x_1(\tau) = A_1\,\cos\tau + B_1\,\sin\tau + \frac{1}{32}\left(3\,C_1^2 - C_2^2\right)\,C_2\,\cos 3\tau - \frac{1}{32}\left(C_1^2 - 3\,C_2^2\right)\,C_1\,\sin 3\tau\,. \tag{6.125}$$

Now use the initial conditions given in equation (6.115) (note that $\dfrac{\partial x_1(0)}{\partial \tau} = -\,\omega_1\,C_2 = 0$ since $\omega_1 = 0$) to obtain

$$A_1 = -\,\frac{1}{32}\left(3\,C_1^2 - C_2^2\right)C_2 \qquad \text{and} \qquad B_1 = \frac{3}{32}\left(C_1^2 - 3\,C_2^2\right)C_1\,. \tag{6.126}$$

Hence,

$$\begin{aligned} x_1(\tau) \;\; = \;\; & -\,\frac{1}{32}\left(3\,C_1^2 - C_2^2\right)\,C_2\,\cos\tau + \frac{3}{32}\left(C_1^2 - 3\,C_2^2\right)\,C_1\,\sin\tau \\ & + \frac{1}{32}\left(3\,C_1^2 - C_2^2\right)\,C_2\,\cos 3\tau - \frac{1}{32}\left(C_1^2 - 3\,C_2^2\right)\,C_1\,\sin 3\tau \end{aligned} \tag{6.127}$$

Finally, the first order approximate solution is given by

$$\begin{aligned} x(\tau) \;\; = \;\; & x_0(\tau) + \epsilon\,x_1(\tau) + O(\epsilon^2) \\ = \;\; & C_1\,\cos\tau + C_2\,\sin\tau \\ & - \frac{\epsilon}{32}\left[\left(3\,C_1^2 - C_2^2\right)C_2\,\cos\tau - 3\,\left(C_1^2 - 3\,C_2^2\right)C_2\,\sin\tau\right] \\ & + \frac{\epsilon}{32}\left[\left(3\,C_1^2 - C_2^2\right)C_2\,\cos 3\tau - \left(C_1^2 - 3\,C_2^2\right)C_1\,\sin 3\tau\right] + O(\epsilon^2)\,. \end{aligned} \tag{6.128}$$

If let

$$C_1 = a\,\cos\phi \qquad \text{and} \qquad C_2 = -\,a\,\sin\phi\,, \tag{6.129}$$

that is,

$$a = (C_1^2 + C_2^2)^{1/2} \qquad \text{and} \qquad \phi = -\tan^{-1}\left(\frac{C_2}{C_1}\right) \quad (\text{or} \quad \tan(\phi) = -\frac{C_2}{C_1}). \tag{6.130}$$

Then,

$$\begin{aligned}
\left(3\,C_1^2 - C_2^2\right) C_2 &= \left(3\,a^2\,\cos^2\phi - a^2\,\sin\phi\right)(-\,a\,\sin\phi) \\
&= a^3\left(3\,\sin\phi - 4\,\sin^3\phi\right) \\
&= -\,a^3\,\sin 3\phi \\
\left(C_1^2 - 3\,C_2^2\right) C_1 &= \left(a^2\,\cos^2\phi - 3\,a^2\,\sin\phi\right) a\,\cos\phi \\
&= a^3\left(4\,\cos^3\phi - 3\,\sin\phi\right) \\
&= a^3\,\cos 3\phi
\end{aligned} \tag{6.131}$$

and thus

$$\begin{aligned}
x_0(\tau) &= a\,\cos(\tau+\phi) \\
x_1(\tau) &= \frac{1}{32}\,a^3\,\sin 3\phi\,\cos\tau + \frac{3}{32}\,a^3\,\cos 3\phi\,\sin\tau - \frac{1}{32}\,a^3\,\sin 3\phi\,\cos 3\tau - \frac{1}{32}\,a^3\,\cos 3\phi\,\sin 3\tau \\
&= \frac{1}{32}\,a^3\,\sin 3\phi\,\cos\tau + \frac{3}{32}\,a^3\,\cos 3\phi\,\sin\tau - \frac{1}{32}\,a^3\,\sin 3(\tau+\phi)\,.
\end{aligned} \tag{6.132}$$

Finally,

$$\begin{aligned}
x(\tau) &= a\,\cos(\tau+\phi) + \frac{\epsilon}{32}\,a^3\,(\sin 3\phi\,\cos\tau + 3\,\cos 3\phi\,\sin\tau) - \frac{\epsilon}{32}\,a^3\,\sin 3(\tau+\phi) + O(\epsilon^2) \\
&= 2\,\cos(\tau+\phi) + \frac{\epsilon}{4}\,(\sin 3\phi\,\cos\tau + 3\,\cos 3\phi\,\sin\tau) - \frac{\epsilon}{4}\,\sin 3(\tau+\phi) + O(\epsilon^2)
\end{aligned} \tag{6.133}$$

where $a = (C_1^2 + C_2^2)^{1/2} = \sqrt{4} = 2$ has been used. This solution form is simpler and clearly shows the relationship between the phase of lower and higher harmonics. Furthermore, it is shown that the phase of the motion is a constant, in addition to that the amplitude is a constant. In fact, the solution form given in polar coordinates could be used at the very beginning of the analysis to simplify the procedure. To do this, let the solution of the ϵ^0-order perturbation equation (6.113) be

$$x_0(\tau) = a\,\cos(\tau+\phi) \tag{6.134}$$

where $a(>0)$ and ϕ are constants, determined by C_1 and C_2 as

$$a = (C_1^2 + C_2^2)^{1/2} \qquad \text{and} \qquad \phi = -\tan^{-1}(C_2/C_1)\,. \tag{6.135}$$

Then the ϵ^1-order perturbation equation (6.114) becomes

$$\begin{aligned} \frac{\partial^2 x_1}{\partial \tau^2} + x_1 &= 2\,\omega_1\, a\, \cos(\tau+\phi) + \left[a^2\, \cos^2(\tau+\phi) - 1\right] a\, \sin(\tau+\phi) \\ &= 2\,\omega_1\, a\, \cos(\tau+\phi) + \left(\frac{1}{4}\,a^2 - 1\right) a\, \sin(\tau+\phi) + \frac{1}{4}\,a^3\, \sin 3(\tau+\phi) \end{aligned} \tag{6.136}$$

which leads to

$$\omega_1 = 0 \qquad \text{and} \qquad \frac{1}{4}\,a^2 - 1 = 0 \;\; \text{i.e.,} \;\; a = 2 \tag{6.137}$$

and so

$$\frac{\partial^2 x_1}{\partial \tau^2} + x_1 = 2\, \sin 3(\tau+\phi) \tag{6.138}$$

which, in turn, yields the solution

$$x_1(\tau) = a_1\, \cos(\tau+\phi_1) - \frac{1}{4}\, \sin 3(\tau+\phi)\,. \tag{6.139}$$

Then from the initial conditions $x_1(0) = \dfrac{\partial x_1(0)}{\partial \tau} = 0$, one obtains $a_1\, \cos\phi_1 = \dfrac{1}{4}\, \sin 3\phi$ and $a_1\, \sin\phi_1 = -\dfrac{3}{4}\, \cos 3\phi$. Hence,

$$x_1(\tau) = \frac{1}{4}\, \sin 3\phi\, \cos\tau + \frac{3}{4}\, \cos 3\phi\, \sin\tau - \frac{1}{4}\, \sin 3(\tau+\phi)\,. \tag{6.140}$$

Therefore, the first-order approximate solution is given by

$$x(\tau) = 2\, \cos(\tau+\phi) + \frac{\epsilon}{4}\, (\sin 3\phi\, \cos\tau + 3\, \cos 3\phi\, \sin\tau) - \frac{\epsilon}{4}\, \sin 3(\tau+\phi) + O(\epsilon^2) \tag{6.141}$$

which is identical to what found above.

6.6.3 Time averaging. This example has been analyzed by using the time averaging in Section 6.4, where the steady-state is found as $a = 2$ which agrees with that obtained from Lindstedt-Poincaré method. Next, we will apply the multiple time scales to consider this example.

6.6.4 Multiple time scales. Use the formulae given in Section 6.5, we have

$$\epsilon^0 \; : \quad D_0^2\, x_0 + x_0 = 0 \tag{6.142}$$

$$\epsilon^1 \; : \quad D_0^2\, x_1 + x_1 = -2\, D_1\, D_0\, x_0 - (x_0^2 - 1)\, D_0\, x_0 \tag{6.143}$$

etc.

The solution for the ϵ^0-order equation (6.142 is

$$x_0 = a(T_1)\, \cos[\, T_0 + \phi(T_1)\,] \tag{6.144}$$

where the amplitude a and the phase ϕ are expressed as functions of T_1 only because we are looking for solutions up to $O(\epsilon)$ order. Then by using this solution, the ϵ^1-order equation becomes

$$\begin{aligned}
& D_0^2\, x_1 + x_1 \\
= \; & -2\, D_1\, [-\, a\, \sin(T_0 + \phi)] - \left[a^2 \cos^2(T_0 + \phi) - 1\right] [-\, a \sin(T_0 + \phi)] \\
= \; & 2\, D_1 a\, \sin(T_0 + \phi) + 2\, a\, D_1\phi\, \cos(T_0 + \phi) + a\, \sin(T_0 + \phi) \left[\frac{a^2}{2} - 1 + \frac{a^2}{2} \cos 2(T_0 + \phi)\right] \\
= \; & \left[2\, D_1 a - a \left(1 - \frac{1}{4} a^2\right)\right] \sin(T_0 + \phi) + 2\, a\, D_1\phi\, \cos(T_0 + \phi) + \frac{1}{4}\, a^3 \sin 3(T_0 + \phi)\,.
\end{aligned} \tag{6.145}$$

Eliminating the secular terms, which may appear in solution x_1, requires

$$\begin{aligned}
D_1 a &= \frac{1}{2}\, a \left(1 - \frac{1}{4}\, a^2\right) \\
D_1\phi &= 0
\end{aligned} \tag{6.146}$$

and thus

$$D_0^2\, x_1 + x_1 = \frac{1}{4}\, a^3\, \sin 3(T_0 + \phi) \tag{6.147}$$

which yields the solution

$$x_1 = -\, \frac{1}{32}\, a^3\, \sin 3(T_0 + \phi)\,. \tag{6.148}$$

Thus, the approximate solution is given by

$$x(t, \epsilon) = a\, \cos(t + \phi) - \frac{\epsilon}{32} a^3\, \sin 3(t + \phi)\,. \tag{6.149}$$

The amplitude a and the phase ϕ of the above periodic solution can be determined from the following *rate equations* up to $O(\epsilon)$ order:

$$\frac{da}{dt} = \frac{\partial a}{\partial T_1}\frac{\partial T_1}{\partial t} + O(\epsilon^2) = \epsilon\, D_1 a + O(\epsilon^2) \approx \frac{\epsilon}{2}\, a\, \left(1 - \frac{1}{4}\, a^2\right) \tag{6.150}$$

and

$$\frac{d\phi}{dt} = \frac{\partial \phi}{\partial T_1}\frac{\partial T_1}{\partial t} + O(\epsilon^3) = \epsilon\, D_1 \phi + O(\epsilon^2) \approx 0\,. \tag{6.151}$$

The above two equations are identical to that obtained by using time averaging method. Setting $\dfrac{da}{dt} = 0$ gives $\bar{a} = 2$ and (6.151) yields $\bar{\phi} = \phi_0$, and thus the steady-state solution is

$$x(t) = 2\,\cos(t + \phi_0) - \frac{\epsilon}{4}\,\sin 3(t + \phi_0)\,. \tag{6.152}$$

To perform the stability analysis, one needs the averaged equation $\dfrac{da}{dt}$ only. Since the derivative of $\dfrac{da}{dt}$ with respect to a is equal to $\epsilon\left(\dfrac{1}{2} - \dfrac{3}{8}\, a^2\right)$ which, in turn, gives $-\epsilon$ for the steady-state solution (i.e., when $a = \bar{a} = 2$). Therefore, the steady state solution is stable (here it is assumed that $\epsilon > 0$).

It should be noted that the first order approximate solution given by the time averaging does not include the ϵ-order correction term $-\dfrac{\epsilon}{4}\,\sin 3(t + \phi_0)$ and so the approximate solution given by the averaging method is less accurate than that given by multiple time scales and Lindstedt-Poincaré procedure as well. However, the process of the time averaging to find it's first order approximations seems being the simplest and it also provides the rate equations for stability analysis. The Lindstedt-Poincaré procedure cannot be used for stability analysis though it produces the same accurate approximations as that of using multiple time scales. So, comparing the three approaches, the multiple time scales is the best one because it not only provides stability analysis but also gives relatively accurate steady-state solutions.

To show that the multiple time scales (and time averaging as well) can produce transient solutions, integrate the equation $\dfrac{da}{dt}$ to obtain

$$\frac{da}{a\,(1 - a^2/4)} = \frac{\epsilon}{2}\, dt \tag{6.153}$$

$$\Longrightarrow$$

$$\int \left[\frac{1}{a} + \frac{-1/4}{1+1/2\,a} + \frac{1/4}{1-1/2\,a}\right] da = \int \frac{\epsilon}{2}\, dt$$

$$\Longrightarrow$$

$$\ln \frac{a}{\sqrt{1-1/4\,a^2}} = \frac{\epsilon}{2}\,(t+\bar{C})$$

$$\Longrightarrow$$

$$\frac{a}{\sqrt{1-1/4\,a^2}} = C\, e^{\epsilon/2\,t} \quad (C = e^{\epsilon/2\,\bar{C}} > 0)$$

$$\Longrightarrow$$

$$a^2 = \frac{C^2\, e^{\epsilon\, t}}{1 + C^2\, e^{\epsilon\, t}/4} = \frac{4}{1 + 4\, C^{-2}\, e^{-\epsilon\, t}}\,.$$

Thus,

$$a = 2\,\left(1 + \frac{4}{C^2\, e^{\epsilon\, t}}\right)^{-\,1/2}. \tag{6.154}$$

Suppose that $a(0) = a_0$ when $t = 0$, then

$$a_0 = 2\,\left(1 + \frac{4}{C^2}\right)^{-\,1/2}, \tag{6.155}$$

so

$$C^2 = \frac{4\, a_0^2}{4 - a_0^2}\,. \tag{6.156}$$

Substituting the above equation into a yields

$$a = 2\,\left[1 - \left(1 - \frac{4}{a_0^2}\right) e^{-\,\epsilon\, t}\right]^{-\,1/2} \tag{6.157}$$

which represents the amplitude of a transient solution for an arbitrary initial condition a_0. Note that $a \to 2$ as $t \to \infty$, as expected.

6.7 An Alternative procedure for Poincaré-Lindstedt Method

Here we introduce an alternate procedure for Poincaré-Lindstedt method, which is easier to be used. You may have noticed in the previous section that when we compare the different perturbation methods using the Van der Pol equation the asymptotic approximate solution

(6.133) obtained through the Poincaré-Lindstedt procedure is different from that derived from the multiple scales (see equation (6.152)). We write down the equations (6.133) for convenience:

$$x(t) = 2\,\cos(t+\phi) + \frac{\epsilon}{4}\,(\sin 3\phi\,\cos t + 3\,\cos 3\phi\,\sin t) - \frac{\epsilon}{4}\,\sin 3(t+\phi) + O(\epsilon^2) \tag{6.158}$$

for the Poincarë-Lindstedt method, where τ has been replaced by t since $\tau \approx (1+O(\epsilon))\,t$; and

$$x(t) = 2\,\cos(t+\phi_0) - \frac{\epsilon}{4}\,\sin 3(t+\phi_0) \tag{6.159}$$

for the multiple scales. It can be seen that these two equations are not identical and equation (6.158) has an additional term $\frac{\epsilon}{4}\,(\sin 3\phi\,\cos\tau + 3\,\cos 3\phi\,\sin\tau)$.

Why are they different? Or are they still equivalent?

This difference comes from the way we deal with the initial conditions for those ordered perturbations in the method of Poincaré-Lindstedt procedure. Recall the approach we used to find the initial conditions: Suppose we use the Poincaré-Lindstedt approach to consider the following differential equation

$$\ddot{x} + \omega_0^2\,x = \epsilon\,f(x,\,\dot{x};\,\epsilon) \tag{6.160}$$

with initial conditions

$$x(0) = C_1 \qquad \text{and} \qquad \dot{x}(0) = C_2\,. \tag{6.161}$$

Here, ω_0 is the frequency of the corresponding free vibration system (i.e. $\epsilon = 0$). $\omega_0 = 1$ for the Van der Pol's equation. In fact, any equation with $\omega_0 \neq 1$ can be transformed into a system with $\omega_0 = 1$ by a time scaling.

By using the Poincaré-Lindstedt method, we expand the frequency $\omega(\epsilon)$ and solution $x(\tau,\,\epsilon)$ as

$$\omega(\epsilon) = \omega_0 + \epsilon\,\omega_1 + \epsilon^2\,\omega_2 + \epsilon^3\,\omega_3 + \cdots \tag{6.162}$$

and

$$x(\tau,\,\epsilon) = x_0(\tau) + \epsilon\,x_1(\tau) + \epsilon^2\,x_2(\tau) + \epsilon^3\,x_3(\tau) + \cdots \tag{6.163}$$

respectively, where

$$\tau = \omega(\epsilon)\,t\,. \tag{6.164}$$

Thus,

$$\frac{dx(t)}{dt} = \omega(\epsilon)\,\frac{\partial x(\tau)}{\partial \tau} \qquad \text{and} \qquad \frac{d^2x(t)}{dt^2} = \omega^2(\epsilon)\,\frac{\partial^2 x(\tau)}{\partial \tau^2}\,. \tag{6.165}$$

Therefore, the initial conditions $x(0) = C_1$ and $\dfrac{dx(0)}{dt} = C_2$ of the solution $x(t)$ can be transformed to the initial conditions of the solution $x(\tau)$ as follows.

Since $t = 0 \Longleftrightarrow \tau = 0$, so $x(t=0) \Longleftrightarrow x(\tau = 0)$ which yields

$$x(t)|_{t=0} = C_1 = x(\tau)|_{\tau=0} = x_0(0) + \epsilon\, x_1(0) + \epsilon^2\, x_2(0) + \epsilon^3\, x_3(0) + \cdots\,. \tag{6.166}$$

By balancing the coefficients of like powers of ϵ, we have

$$x_0(0) = C_1 \qquad \text{and} \qquad x_i(0) = 0 \quad \forall \;\; i \geq 1\,. \tag{6.167}$$

Similarly, by balancing the first derivatives we may find

$$\begin{aligned}
&\frac{dx(t)}{dt}|_{t=0} = C_2 = \omega(\epsilon)\,\frac{\partial x(\tau)}{\partial \tau}|_{\tau=0} \\
= \;& \left(\omega_0 + \epsilon\,\omega_1 + \epsilon^2\,\omega_2 + \epsilon^3\,\omega_3 + \cdots\right)\left(\frac{\partial x_0(0)}{\partial \tau} + \epsilon\,\frac{\partial x_1(0)}{\partial \tau} + \epsilon^2\,\frac{\partial x_2(0)}{\partial \tau} + \epsilon^3\,\frac{\partial x_3(0)}{\partial \tau} + \cdots\right) \\
= \;& \omega_0\,\frac{\partial x_0(0)}{\partial \tau} + \epsilon\left(\omega_0\,\frac{\partial x_1(0)}{\partial \tau} + \omega_1\,\frac{\partial x_0(0)}{\partial \tau}\right) \\
&\quad + \epsilon^2\left(\omega_0\,\frac{\partial x_2(0)}{\partial \tau} + \omega_1\,\frac{\partial x_1(0)}{\partial \tau} + \omega_2\,\frac{\partial x_0(0)}{\partial \tau}\right) \\
&\quad + \epsilon^3\left(\omega_0\,\frac{\partial x_3(0)}{\partial \tau} + \omega_1\,\frac{\partial x_2(0)}{\partial \tau} + \omega_2\,\frac{\partial x_1(0)}{\partial \tau} + \omega_3\,\frac{\partial x_0(0)}{\partial \tau}\right) + \cdots
\end{aligned} \tag{6.168}$$

which results in, by balancing the coefficients of ϵ

$$\begin{aligned}
\frac{\partial x_0(0)}{\partial \tau} &= \left(\frac{1}{\omega_0}\right) C_2 \\
\frac{\partial x_1(0)}{\partial \tau} &= -\left(\frac{\omega_1}{\omega_0}\right)\frac{\partial x_0(0)}{\partial \tau} = -\left(\frac{\omega_1}{\omega_0^2}\right) C_2 \\
\frac{\partial x_2(0)}{\partial \tau} &= -\,\frac{1}{\omega_0}\left(\omega_1\,\frac{\partial x_1(0)}{\partial \tau} + \omega_2\,\frac{\partial x_0(0)}{\partial \tau}\right) = \left(\frac{-\,\omega_0\,\omega_2 + \omega_1^2}{\omega_0^3}\right) C_2 \\
\frac{\partial x_3(0)}{\partial \tau} &= -\,\frac{1}{\omega_0}\left(\omega_1\,\frac{\partial x_2(0)}{\partial \tau} + \omega_2\,\frac{\partial x_1(0)}{\partial \tau} + \omega_3\,\frac{\partial x_0(0)}{\partial \tau}\right) \\
&= \left(\frac{-\,\omega_0^2\,\omega_3 + 2\,\omega_0\,\omega_1\,\omega_2 - \omega_1^3}{\omega_0^4}\right) C_2 \\
&\;\;\vdots
\end{aligned} \tag{6.169}$$

An alternate way is that let these initial conditions not determined at this stage and also let the coefficients in the solutions of $x_0(\tau)$, $x_1(\tau)$, etc. undetermined until we find the complete approximate solution $x(\tau)$ at the end of analysis.

To demonstrate this procedure, let's consider the Van der Pol's equation

$$\ddot{x} + x + \epsilon\,(x^2 - 1)\,\dot{x} = 0\,. \tag{6.170}$$

This nonlinear DE can be transformed to a set of DEs by using the equations (6.162)–(6.165) as follows:

$$\epsilon^0 \;:\quad \frac{\partial^2 x_0}{\partial \tau^2} + x_0 = 0 \tag{6.171}$$

$$\epsilon^1 \;:\quad \frac{\partial^2 x_1}{\partial \tau^2} + x_1 = -\,2\,\omega_1\,\frac{\partial^2 x_0}{\partial \tau^2} - (x_0^2 - 1)\,\frac{\partial x_0}{\partial \tau} \tag{6.172}$$

etc.

From the zero-order equation (6.171), we obtain the solution

$$x_0(\tau) = A_0\,\cos\tau + B_0\,\sin\tau \tag{6.173}$$

where A_0 and B_0 are arbitrary coefficients. With this solution, the first-order equation (6.172) becomes

$$\begin{aligned}
\frac{\partial^2 x_1}{\partial \tau^2} + x_1 \;&=\; -\,2\,\omega_1\,(-\,A_0\,\cos\tau - B_0\,\sin\tau) \\
&\quad -\,(A_0^2\,\cos^2\tau + B_0^2\,\sin^2\tau + 2\,A_0\,B_0\,\sin\tau\,\cos\tau - 1)\,(-\,A_0\,\sin\tau + B_0\,\cos\tau) \\
&=\; \cos\tau\left\{2\,\omega_1\,A_0 + B_0\,\left[1 - \frac{1}{4}\,(A_0^2 + B_0)\right]\right\} \\
&\quad +\,\sin\tau\left\{2\,\omega_1\,B_0 - A_0\,\left[1 - \frac{1}{4}\,(A_0^2 + B_0)\right]\right\} \\
&\quad -\,\frac{1}{4}\,\left(3\,A_0^2 - B_0^2\right)\,B_0\,\cos 3\tau + \frac{1}{4}\,\left(A_0^2 - 3\,B_0^2\right)\,A_0\,\sin 3\tau\,.
\end{aligned} \tag{6.174}$$

Therefore,

$$\begin{aligned}
2\,\omega_1\,A_0 + B_0\,\left[1 - \frac{1}{4}\,(A_0^2 + B_0)\right] \;&=\; 0 \\
2\,\omega_1\,B_0 - A_0\,\left[1 - \frac{1}{4}\,(A_0^2 + B_0)\right] \;&=\; 0
\end{aligned} \tag{6.175}$$

which yields

$$\omega_1 = 0\,, \tag{6.176}$$

and

$$A_0^2 + B_0^2 = 0 \qquad \text{or} \qquad A_0^2 + B_0^2 = 4\,. \tag{6.177}$$

We take the nontrivial solution $A_0^2 + B_0^2 = 4$ which represents steady-state periodic solutions (limit cycles). Now equation (6.174) becomes

$$\frac{\partial^2 x_1}{\partial \tau^2} + x_1 = -\frac{1}{4}\left(3\,A_0^2 - B_0^2\right) B_0 \cos 3\tau + \frac{1}{4}\left(A_0^2 - 3\,B_0^2\right) A_0 \sin 3\tau \tag{6.178}$$

from which one finds the solution

$$x_1(\tau) = A_1 \cos\tau + B_1 \sin\tau + \frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0 \cos 3\tau - \frac{1}{32}\left(A_0^2 - 3\,B_0^2\right) A_0 \sin 3\tau \tag{6.179}$$

where A_1 and B_1 are arbitrary coefficients.

The first-order approximate solution can now written as

$$\begin{aligned} x(\tau) &= x_0(\tau) + \epsilon\, x_1(\tau) + O(\epsilon^2) \\ &= A_0 \cos\tau + B_0 \sin\tau \\ &\quad + \epsilon\,\Big[\, A_1 \cos\tau + B_1 \sin\tau \\ &\qquad + \frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0 \cos 3\tau - \frac{1}{32}\left(A_0^2 - 3\,B_0^2\right) A_0 \sin 3\tau \Big] + O(\epsilon^2)\,. \end{aligned} \tag{6.180}$$

and

$$\omega = 1 + \epsilon\,\omega_1 + O(\epsilon^2) = 1 + O(\epsilon^2) \qquad \Longrightarrow \qquad \tau = \omega(\epsilon)\, t \approx t\,. \tag{6.181}$$

So, we have

$$\begin{aligned} x(t) &= A_0 \cos t + B_0 \sin t \\ &\quad + \epsilon\,\Big[\, A_1 \cos t + B_1 \sin t \\ &\qquad + \frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0 \cos 3t - \frac{1}{32}\left(A_0^2 - 3\,B_0^2\right) A_0 \sin 3t \Big] + O(\epsilon^2)\,. \end{aligned} \tag{6.182}$$

Next, from the initial conditions: $x(0) = C_1$ and $\dot{x}(0) = C_2$, we obtain

$$\begin{aligned} x(0) &= C_1 = A_0 + \epsilon\left[A_1 + \frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0\right] \\ \dot{x}(0) &= C_2 = B_0 + \epsilon\left[B_1 - \frac{3}{32}\left(A_0^2 - 3\,B_0^2\right) A_0\right] \end{aligned}$$

or

$$\begin{aligned} A_0 + \frac{\epsilon}{32}\left(3\,A_0^2 - B_0^2\right) B_0 + \epsilon\, A_1 &= C_1 \\ B_0 - \frac{3\,\epsilon}{32}\left(A_0^2 - 3\,B_0^2\right) A_0 + \epsilon\, B_1 &= C_2\,. \end{aligned} \tag{6.183}$$

The two equations given in (6.183) have four unknowns A_0, B_0, A_1 and B_1 (more unknowns will be involved if go to higher order perturbations) and therefore, they may have infinite solutions. For example, $A_0 = C_1$, $B_0 = C_2$, $A_1 = -\frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0$, $B_1 = \frac{3}{32}\left(A_0^2 - 3\,B_0^2\right) A_0$ is just one of them.

An alternate solution, which may be more convenient to be applied in perturbation analysis, is that

$$A_1 = B_1 = 0\,, \tag{6.184}$$

and A_0 and B_0 are determined from the following equations:

$$\begin{aligned} A_0 + \frac{\epsilon}{32}\left(3\,A_0^2 - B_0^2\right) B_0 &= C_1 \\ B_0 - \frac{3\,\epsilon}{32}\left(A_0^2 - 3\,B_0^2\right) A_0 &= C_2\,. \end{aligned} \tag{6.185}$$

Since for this Van der Pol's equation, A_0 and B_0 must satisfy $A_0^2 + B_0^2 = 4$, so C_1 and C_2 must also satisfy a similar condition. Suppose we are only interested in steady-state solutions (in fact, only steady-state solutions are available if using Poincaré-Lindstedt method), then we even do not need to solve A_0 and B_0 from C_1 and C_2 because A_0 and B_0 are initial conditions.

Hence, we have the first-order approximate solution

$$\begin{aligned} x(t) &= A_0\,\cos t + B_0\,\sin t \\ &\quad + \frac{\epsilon}{32}\left[\left(3\,A_0^2 - B_0^2\right) B_0\,\cos 3t - \left(A_0^2 - 3\,B_0^2\right) A_0\,\sin 3t\right] + O(\epsilon^2) \end{aligned} \tag{6.186}$$

with $A_0^2 + B_0^2 = 4$. Now by using the transformation

$$\begin{aligned} a &= \sqrt{A_0^2 + B_0^2} = 2 \\ \phi_0 &= -\tan^{-1}\left(\frac{B_0}{A_0}\right), \end{aligned} \tag{6.187}$$

one finally obtains

$$x(t) = 2\,\cos(t+\phi_0) - \frac{\epsilon}{4}\,\sin 3(t+\phi_0) + O(\epsilon^2) \tag{6.188}$$

which is identical to equation (6.159) or (6.152) obtained using the multiple scales.

This alternate procedure for Poincaré-Lindstedt method is easier to be used than the first procedure. The idea used in this approach is, unlike the first approach to divide the initial conditions into all the ordered perturbation equations, to include the initial conditions in just zero-order perturbation equation only. By this approach, we do not need to derive the solution for the homogeneous part, in other words, we can set $A_i = B_i = 0\ \forall\ i \geq 1$, and only need to find the particular solution for the non-homogeneous part. In deed, this procedure is used in most perturbation techniques such as multiple scales and harmonic balancing in finding asymptotic periodic or quasiperiodic solutions.

Finally, we answer the question raised at the beginning of this section: do the two procedures give the same results? In other words, are the solutions given in (6.158) and (6.159) equivalent? The answer is **yes**! This becomes clear if we consider their original forms

$$\begin{aligned} x(t) \;=\; & C_1\,\cos t + C_2\,\sin t \\ & + \epsilon\,\Big[-\frac{1}{32}\left(3\,C_1^2 - C_2^2\right) C_2\,\cos t + \frac{3}{32}\,\left(C_1^2 - 3\,C_2^2\right) C_2\,\sin t \\ & \quad + \frac{1}{32}\left(3\,C_1^2 - C_2^2\right) C_2\,\cos 3t - \frac{1}{32}\left(C_1^2 - 3\,C_2^2\right) C_1\,\sin 3t \Big] + O(\epsilon^2) \end{aligned} \tag{6.189}$$

and

$$\begin{aligned} x(t) \;=\; & A_0\,\cos t + B_0\,\sin t \\ & + \epsilon\,[\,A_1\,\cos t + B_1\,\sin t \\ & \quad + \frac{1}{32}\left(3\,A_0^2 - B_0^2\right) B_0\,\cos 3t - \frac{1}{32}\left(A_0^2 - 3\,B_0^2\right) A_0\,\sin 3t \Big] + O(\epsilon^2) \end{aligned} \tag{6.190}$$

where the coefficients C_1, C_2 and A_0, B_0 are related by equation (6.183). Note that here the only constraint is $A_0^2 + B_0^2 = 4$. Only if choosing $A_0 = C_1$, $B_0 = C_2$, $A_1 = -\frac{1}{32}\left(3\,C_1^2 - C_2^2\right) C_2$ and $B_1 = \frac{3}{32}\left(C_1^2 - 3\,C_2^2\right) C_2$, then $C_1^2 + C_2^2 = 4$. Substituting (6.183) into (6.189) results in (6.190), which indicates that (6.189) and (6.190), or equivalently (6.158) and (6.159), are indeed identical via the transformation (6.183).

Chapter 7

EXISTENCE AND UNIQUENESS THEOREM

For this chapter, we need to adopt a different perspective. If we want to watch the "dynamics" of a system, we need to keep "track" of everything that is going on and not just the "solution". The fact of the matter is that we know that in mechanics we need to keep track of position and momentum in order to keep full contact with the dynamics.

In this regard there is a very general theorem relating to differential equations to systems of differential equations.

7.1 Systems of Equations

Given a differential equation of order n, which for simplicity we shall suppose solved for the derivative of highest order; i.e., we shall suppose that

$$y^{(n)} = f\left(x,\, y,\, y',\, \cdots,\, y^{(n-1)}\right) \tag{7.1}$$

where $y' = \dfrac{dy}{dx}$, $y'' = \dfrac{d^2y}{dx^2}$, $y^{(k)} = \dfrac{d^ky}{dx^k}$, etc. If we set

$$\begin{aligned} y &= y_1 \\ y' &= y_2 \\ y'' &= y_3 \\ &\vdots \\ y^{(n-1)} &= y_n, \end{aligned} \tag{7.2}$$

then, we have a system of n differential equations of first order:

$$\begin{aligned} {y_1}' &= y_2 \\ {y_2}' &= y_3 \\ {y_3}' &= y_4 \end{aligned}$$

$$\begin{aligned} &\vdots \\ y'_{(n-1)} &= y_n \\ y_n' &= f(x,\, y_1,\, y_2,\, \cdots,\, y_n). \end{aligned} \tag{7.3}$$

This is a particular case of a system of the following differential equations:

$$\begin{aligned} y_1' &= f_1(x,\, y_1,\, y_2,\, \cdots,\, y_n) \\ y_2' &= f_2(x,\, y_1,\, y_2,\, \cdots,\, y_n) \\ &\vdots \\ y_n' &= f_n(x,\, y_1,\, y_2,\, \cdots,\, y_n). \end{aligned} \tag{7.4}$$

or

$$y_i' = f_i(x,\, y_1,\, y_2,\, \cdots,\, y_n) \qquad i = 1,\, 2,\, 3,\, \ldots \tag{7.5}$$

To give ourselves some notational relief, we will later adopt the customary (dynamical systems) approach to such a system of DEs

$$\boldsymbol{y}' = \boldsymbol{f}(x,\, \boldsymbol{y}) \tag{7.6}$$

where $\boldsymbol{y}$, $\boldsymbol{y}'$ and $\boldsymbol{f}$ are vector functions with n components.

The converse is also true. If we differentiate each equation in the system $(n-1)$ times with respect to x, leaving us with n^2 equations all together from which we can extract an equation for only one variable of y, of order n, all other variables being eliminated.

For an initial value problem (IVP) for the original equation, one needs

$$y(0) = y_0\,,\ y'(0) = y_0', \ldots,\, y^{n-1}(0) = y_0^{(n-1)}.$$

This amounts a condition on the column vector $\boldsymbol{y}$ in the revised notation.

Note: This is not so different from the devices used to solve other nonlinear DEs.

EXAMPLE 7.1. Write

$$2\,y''' - 6\,y'' + 4\,y' + y^2 = \sin t \tag{7.7}$$

as a first-order system of differential equations.

Solution: Let

$$y = y_1\,, \quad y' = y_2\,, \quad y'' = y_3\,,$$

then,

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ y_3' &= -\frac{1}{2}\,y_1^2 - 2\,y_2 + 3\,y_3 + \frac{1}{2}\,\sin t. \end{aligned} \tag{7.8}$$

Hence,

$$\frac{d\boldsymbol{y}}{dt} = \frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} y_2 \\ y_3 \\ \frac{1}{2}\,y_1^2 - 2\,y_2 + 3\,y_3 + \frac{1}{2}\,\sin t \end{pmatrix} = \boldsymbol{f}(t,\,\boldsymbol{y}). \tag{7.9}$$

EXAMPLE 7.2. Consider the pendulum example (see equation (6.6))

$$\ddot{\theta} + \omega_0^2\,\sin\theta = 0 \qquad \text{where} \quad \omega_0^2 = \frac{g}{l}. \tag{7.10}$$

Write it as a first-order system of differential equations.

Solution: Let

$$\theta = y_1,\ \dot{\theta} = y_2,$$

then,

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 \\ \frac{dy_2}{dt} &= \ddot{\theta} = -\,\omega_0^2\,\sin y_1 \end{aligned} \tag{7.11}$$

or

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\,\omega_0^2\,\sin y_1 \end{pmatrix}. \tag{7.12}$$

EXAMPLE 7.3. Write the system

$$\begin{aligned} \frac{dx}{dt} &= x - y \\ \frac{dy}{dt} &= 2\,x - y + t \end{aligned} \tag{7.13}$$

as a single 2nd-order differential equation.

Solution: Equation (7.13) can be written as

$$\left(\frac{d}{dt} - 1\right) x + y = 0 \tag{7.14}$$

$$-2\,x + \left(\frac{d}{dt} + 1\right) y = t. \tag{7.15}$$

Then, multiplying (7.14) by $\left(\frac{d}{dt} + 1\right)$ and subtracting (7.15) from the resulting equation yield

$$\left(\frac{d}{dt} + 1\right)\left(\frac{d}{dt} - 1\right) x + 2\,x = -\,t$$

$$\Longrightarrow$$

$$\left(\frac{d}{dt} + 1\right)\left(\frac{dx}{dt} - x\right) + 2\,x = -\,t$$

$$\Longrightarrow$$

$$\frac{d}{dt}\left(\frac{dx}{dt} - x\right) + \left(\frac{dx}{dt} - x\right) + 2\,x = -\,t$$

$$\Longrightarrow$$

$$\frac{d^2x}{dt^2} - \frac{dx}{dt} + \frac{dx}{dt} + x = -\,t$$

and finally we have

$$\frac{d^2x}{dt^2} + x = -\,t \tag{7.16}$$

which is indeed a 2nd-order DE.

The more convenient way to achieve equation (7.16) is to introduce the differential operator $D = \frac{d}{dt}$. With this operator, we can rewrite (7.13) as

$$Dx = x - y$$

$$Dy = 2\,x - y + t$$

or

$$(D-1)x + y = 0 \tag{7.17}$$

$$-2\,x + (D+1)\,y = t \tag{7.18}$$

which can be treated as algebraic equations. Thus, one may eliminate y to obtain an equation of x or eliminate x to get an equation of y.

(I) Equation of x: (7.17) $\times (D+1)-$ (7.18) results in

$$\begin{aligned} & (D+1)(D-1)x + 2x = -t \\ \Longrightarrow \quad & (D^2-1)x + 2x = -t \\ \Longrightarrow \quad & D^2 x - x + 2x = -t \\ \Longrightarrow \quad & D^2 x + x = -t \end{aligned}$$

i.e.,

$$\frac{d^2x}{dt^2} + x = t. \tag{7.19}$$

(II) Equation of y: $2 \times$ (7.17) $+ (D-1) \times$ (7.18) yields

$$\begin{aligned} & 2y + (D-1)(D+1)y = (D-1)t \\ \Longrightarrow \quad & 2y + (D^2-1)y = Dt - t \\ \Longrightarrow \quad & D^2 y + y = 1 - t \end{aligned}$$

or

$$\frac{d^2y}{dt^2} + y = 1 - t. \tag{7.20}$$

Suppose now we want to go back to the original system (7.13) from equation (7.19). Let $y = \dfrac{dx}{dt}$, then,

$$\frac{dy}{dt} = \frac{d^2x}{dt^2} = -x - t.$$

So we have

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -x - t \end{aligned} \tag{7.21}$$

However, this is not the original system (7.13). To go back to the original system from (7.19), we need to set

$$y = x - \frac{dx}{dt}\,, \tag{7.22}$$

then,

$$\begin{aligned}\frac{dy}{dt} &= \frac{dx}{dt} - \frac{d^2x}{dt^2} \\ &= x - y - (-x - t) \\ &= 2\,x - y + t.\end{aligned}$$

Combing the above two equations we have

$$\begin{aligned}\frac{dx}{dt} &= x - y \\ \frac{dy}{dt} &= 2\,x - y + t. \end{aligned} \tag{7.23}$$

Example 7.3 illustrates that the transformation from single equation to a system is not unique. Uniqueness is not the key. Differences do not typically change the facts one discovers from such a strategy.

Why do we want to transfer a single DE to a system of DEs? By doing this, we can apply linear system theory and matrix theory to solve differential equations.

7.2 Phase Space

Describing a system may be not unique, but as we have seen in the previous subsection, a standard form proves to be useful. To identify derivatives up to order $(n-1)$ in an nth order DE with the variables, thus, for a 2nd-order DE we have the variables

$$y \quad \text{and} \quad \frac{dy}{dt} \quad \Longrightarrow \quad (y_1,\, y_2);$$

For a 3rd-order DE,

$$\left(y,\, \frac{dy}{dt},\, \frac{d^2y}{dt^2}\right) \quad \Longrightarrow \quad (y_1,\, y_2,\, y_3);$$

etc., for nth-order DE, we have

$$\left(y,\, \frac{dy}{dt},\, \frac{d^2y}{dt^2},\, \cdots,\, \frac{d^ny}{dt^n}\right) \quad \Longrightarrow \quad (y_1,\, y_2,\, \cdots,\, y_n).$$

It should be clear that we are dealing with a space in which all of the dynamics are being represented simultaneously. The space spanned by $y_1,\, y_2,\, \cdots,\, y_n$ is called *phase space*. For a 2nd-order DE

$$\begin{aligned} y &- \text{position} \longrightarrow y_1 \\ \frac{dy}{dt} &- \text{velocity} \longrightarrow y_2 \end{aligned}$$

and thus the system of equations is given by

$$\begin{aligned} \dot{y}_1 &= f_1(y_1, y_2) \\ \dot{y}_2 &= f_2(y_1, y_2) \end{aligned} \tag{7.24}$$

which is a 2-dimensional phase plane, and it can be put in a more compact form – vector form

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}) \tag{7.25}$$

where $\boldsymbol{y}$ and $\boldsymbol{f}$ are 2-dimensional vectors. In particular, $\boldsymbol{f}(\boldsymbol{y})$ is called *vector field*. Geometrically, $\boldsymbol{y}$ represents a point in the phase space (phase point) and $\dot{\boldsymbol{y}}$ is the velocity vector at that point. See Fig. 7.1, where the curve is usually called *a solution, a trajectory* or *an orbit*.

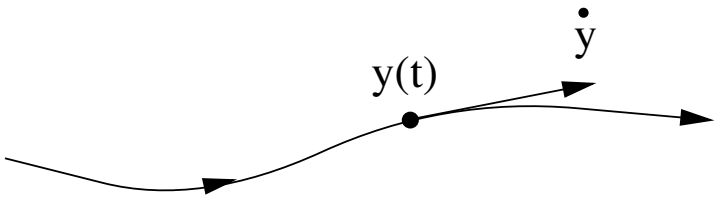


Fig. 7.1 A trajectory.

Consider a simple 2-dim. example

$$\dot{\boldsymbol{y}} = A\,\boldsymbol{y} \qquad \text{where} \qquad A = \begin{bmatrix} 2 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{y} = (y_1 \;\; y_2) \in \boldsymbol{R}^2 \tag{7.26}$$

which is decoupled and can be written as

$$\begin{aligned} \dot{y}_1 &= 2\,y_1 \\ \dot{y}_2 &= -\frac{1}{2}\,y_2\,. \end{aligned} \tag{7.27}$$

The solution of (7.27) is trivial, given by

$$\begin{cases} x_1 = y_{10}\,e^{2t}, \\ x_2 = y_{20}\,e^{-\frac{1}{2}t}. \end{cases} \tag{7.28}$$

From a geometrical viewpoint, regard the solution as a specifying a curve in $\boldsymbol{R}^2$. The initial condition requires the curve to pass through $(y_{10},\, y_{20})$ (see Fig. 7.2).

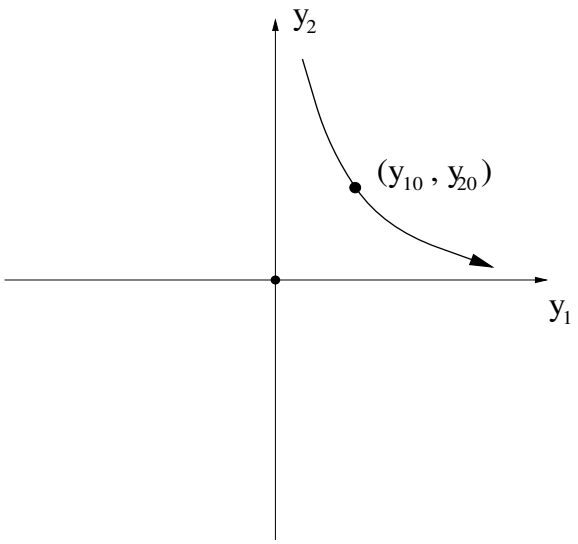


Fig. 7.2 A solution curve in a 2-dim. phase plane.

By sketching this field at various points and pointing them with smooth curves we get a qualitative picture of what the solution looks like (Fig. 7.3). This is called the "phase portrait".

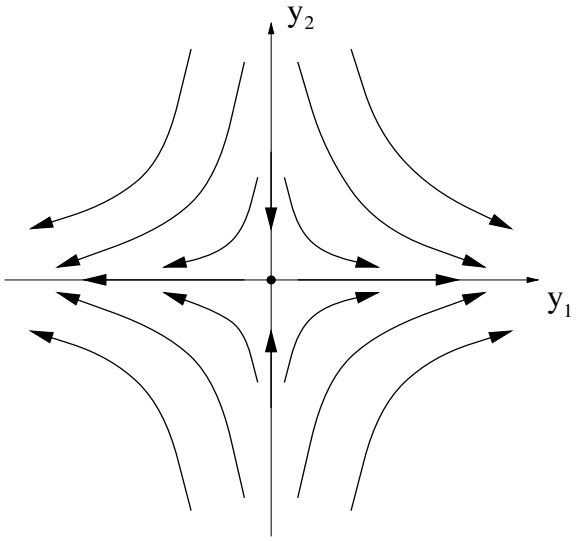


Fig. 7.3 Some solution curves ("phase portrait") to $\dot{\boldsymbol{y}} = A\,\boldsymbol{y}$, $A = \begin{bmatrix} 2 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}$.

For the above simple linear system, we obtained the analytical solution. However, for

nonlinear systems, there is typically no hope of finding the trajectories analytically. Even when explicit formulas are available, they are often too complicated to provide much insight. So, to use the *qualitative* method – phase portrait is very useful. A typical phase portrait of a nonlinear system is given in Fig. 7.4.

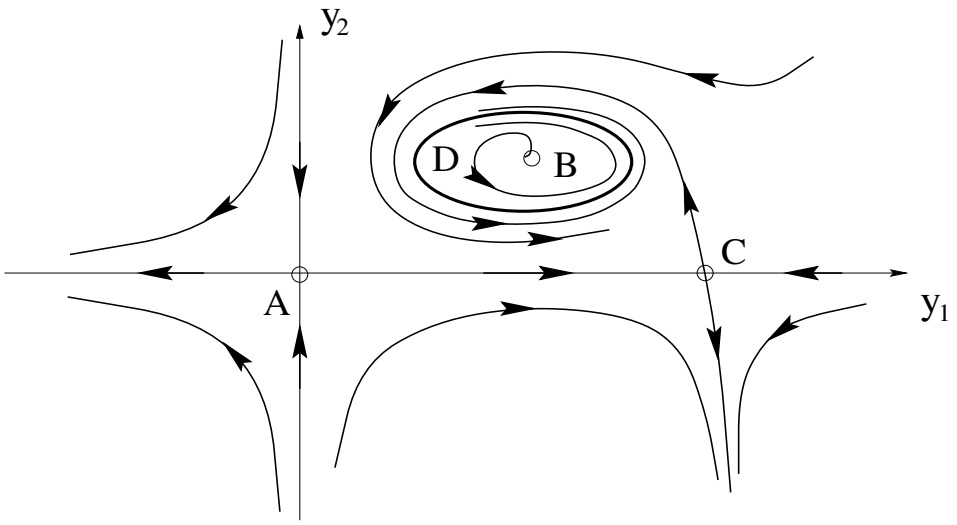


Fig. 7.4 A typical phase of portrait of a nonlinear system.

Most salient features of phase portrait:

1) **Fixed points** A, B, C which are obtained by solving the algebraic equation $\dot{x} = f(x) = 0$ (usually nonlinear equation). They are also called *steady states* or *equilibrium* of the system $\dot{x} = f(x)$.

2) **Closed orbit** D which is periodic motion (or periodic solution).

3) **Pattern** of trajectories near the fixed points and closed orbits.

4) **Stability** or **instability**. For the nonlinear system shown in Fig. 7.4, the fixed points A, B and C are unstable; while the periodic orbit is stable.

For this course, we'll restrict phase portrait to 2-dimensional phase space, i.e., phase plane. To help plot phase portraits one can use so called **nullclines** which are determined from the equation

$$\begin{aligned} \dot{y}_1 &= f_1(y_1,\, y_2) \\ \dot{y}_2 &= f_2(y_1,\, y_2) \end{aligned}$$

by either setting $\dot{y}_1 = 0$ or $\dot{y}_2 = 0$. Consider the following example.

EXAMPLE 7.4. Sketch the phase portrait of the following system

$$\begin{aligned} \dot{y}_1 &= y_1 + e^{-y_2} \\ \dot{y}_2 &= -\,y_2\,. \end{aligned} \tag{7.29}$$

Solution: Setting $\dot{y}_1 = 0$ yields

$$y_1 = -\,e^{-y_2} \tag{7.30}$$

on which $\dot{y}_1 = 0$ and $\dot{y}_2 = -\,y_2$, so the direction of the vector filed is vertical, and points downward when $y_2 > 0$ and upward when $y_2 < 0$.

Similarly, setting $\dot{y}_2 = 0$ results in

$$y_2 = 0 \tag{7.31}$$

which is the y_1-axis. Thus, on this axis, $\dot{y}_2 = 0$ and $\dot{y}_1 = y_1 + e^{-0} = y_2 + 1$. Hence, the vector field points to right direction when $y_1 > -1$ and to the left direction when $y_1 < -1$. These two nullclines are shown in Fig. 7.5. It is seen that y_1–y_2 plane is divided into four regions: A ($\dot{y}_1 > 0$, $\dot{y}_2 < 0$), B ($\dot{y}_1 > 0$, $\dot{y}_2 > 0$), C ($\dot{y}_1 < 0$, $\dot{y}_2 > 0$), and D ($\dot{y}_1 < 0$, $\dot{y}_2 < 0$). Therefore the direction of vector filed can be easily drawn as shown in Fig. 7.5. Smoothly connecting the arrows gives us the phase portrait of system (7.29) in Fig. 7.6.

In general, one needs to use computer to plot a 2-dimensional phase portrait of the system

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y})$$

where $\boldsymbol{y}$ and $\boldsymbol{f}$ are 2-dimensional vectors. Many numerical methods can be used, the following is just an example: *fourth-order Runge-Kutta method.*

$$\begin{aligned} \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \frac{1}{6}\,(\boldsymbol{k}_1 + 2\,\boldsymbol{k}_2 + 2\,\boldsymbol{k}_3 + \boldsymbol{k}_4)\,, \\ \boldsymbol{k}_1 &= \boldsymbol{f}(\boldsymbol{y}_n)\,\Delta t \end{aligned}$$

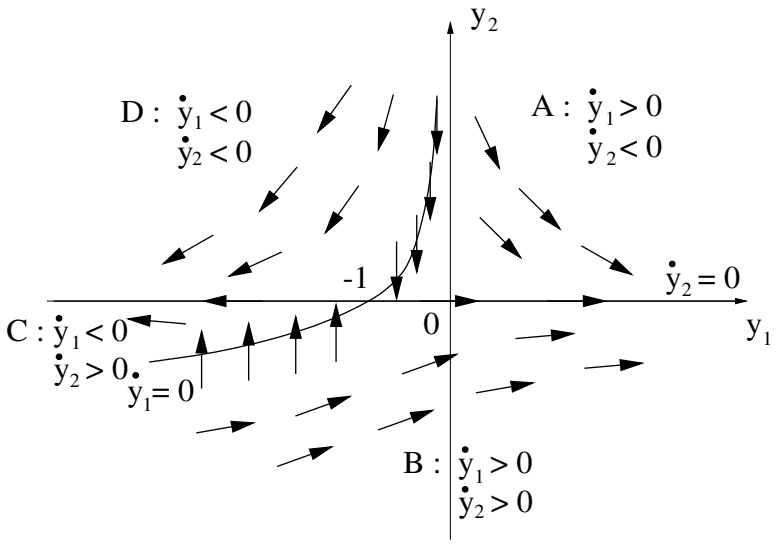


Fig. 7.5 Nullclines with direction of vector filed of system (7.29).

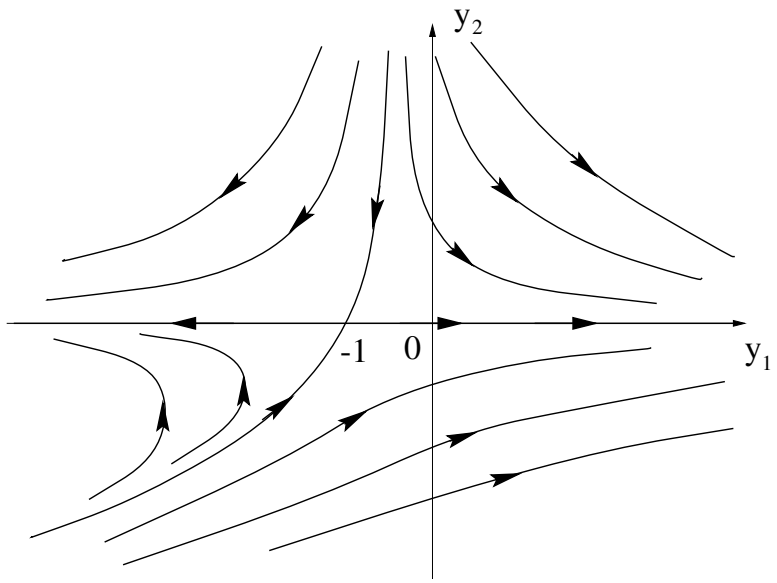


Fig. 7.6 Phase portrait for the system (7.29).

$$\begin{aligned} \boldsymbol{k}_2 &= \boldsymbol{f}(\boldsymbol{y}_n + \frac{1}{2}\boldsymbol{k}_1)\,\Delta t \\ \boldsymbol{k}_3 &= \boldsymbol{f}(\boldsymbol{y}_n + \frac{1}{2}\boldsymbol{k}_2)\,\Delta t \\ \boldsymbol{k}_4 &= \boldsymbol{f}(\boldsymbol{y}_n + \boldsymbol{k}_3)\,\Delta t. \end{aligned} \tag{7.32}$$

Choosing 0.1 as the time step Δt usually provides sufficient accuracy.

7.3 Norms

So far we have discussed a lot about solutions of a differential equation. However, we have no guarantee that a general nonlinear system $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y})$ even *has* solutions!

To discuss the *existence* and *uniqueness* theorem of DEs, we need to define "distance" in phase space. What is the "distance" between two points, or two vectors $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)$ and $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$? We must also consider *distance* or *length* of a vector: *vector norms*. Such things are **not** unique! For example, if we have two vectors $\boldsymbol{a}$ $\boldsymbol{b}$, the "distance" between them depends upon the relative importance of the coordinates: something like "scale" so to speak. Obviously, one can set these things differently. After all, how do you compare velocity to position to define a "distance" in this space?

With a typical *Euclidean norm*, ℓ, we have

$$\ell^2 = |\boldsymbol{a} - \boldsymbol{b}|^2 = \sum_i (a_i - b_i)^2. \tag{7.33}$$

But $\boldsymbol{R}^n$ does not need to be Euclidean in that sense. We could just as well have

$$\ell^m = \sum_i |a_i - b_i|^m. \tag{7.34}$$

We could also have

$$\ell = \sum_i |a_i - b_i| \tag{7.35}$$

which is the so called "Taxi cab" norm because this is the distance a taxi cab travels in moving between two points on a rectangular city road network.

Note: Norms in common language: "How far is that place?", "Is that driving distance (Taxi cab norm) or flying distance (Euclidean norm)?"

We may also have a norm defined by the limit as $m \to \infty$. In this case we write

$$\ell = \left(\sum_i |a_i - b_i|^m \right)^{1/m} = \sup_i \{ |a_i - b_i| \} = |a_k - b_k| \tag{7.36}$$

where the notation sup is called *supremum* which is the *smallest upper bound* of the set $\{\cdots\}$. For example.

$$\sup\{1, 3, 5, 0, -4\} = 5$$

$$\sup\{0.5, 0.6, 0.9, 0.99, 0.999, \cdots\} = 1$$

In our case of the distance definition, the supremum is just the biggest one; i.e., $|a_k - b_k| \geq |a_i - b_i|$ $i = 1, 2, \cdots$ Therefore,

$$\begin{aligned} \ell &= \left(|a_k - b_k|^m \left[1 + \sum_{i,\, i \neq k} \frac{|a_i - b_i|^m}{|a_k - b_k|^m} \right] \right)^{1/m} \\ &= |a_k - b_k| \left(1 + \sum_{i,\, i \neq k} \left[\frac{|a_i - b_i|}{|a_k - b_k|} \right]^m \right)^{1/m} \end{aligned}$$

Now if $|a_k - b_k|$ does not equal any other component's magnitude, then,

$$\left[\frac{|a_i - b_i|}{|a_k - b_k|} \right]^m \rightarrow 0 \quad \text{as} \quad m \rightarrow \infty \quad \forall\, i \neq k.$$

Hence,

$$\ell = |a_k - b_k|. \tag{7.37}$$

If $(a_k - b_k)$ is repeated so that its magnitude appears j times, then,

$$\ell = |a_k - b_k| \left(j + \sum_i \left[\frac{|a_i - b_i|}{|a_k - b_k|} \right]^m \right)^{1/m}$$

where the index of summation, i, goes through all possible choices except those components with magnitude $|a_k - b_k|$. When $m \rightarrow \infty$, ℓ goes to $|a_k - b_k|$ since $\left[\frac{|a_i - b_i|}{|a_k - b_k|} \right]^m \rightarrow 0$ and $j^{1/m} \rightarrow 1$ as $m \rightarrow \infty$. So, in the limit we again have equation (7.37) even when $(a_k - b_k)$ is repeated. This norm is known as ℓ_∞ norm, which is just the magnitude of the largest coordinate.

Each of these norms *implies* the existence of a *distance* function $d(\boldsymbol{a}, \boldsymbol{b})$ where $\boldsymbol{a}$ and $\boldsymbol{b}$ represent n-tuples or "points", or vector in the space. A space with a distance function is known as a *metric space*, because you can "measure it". The properties of a distance function are

(i) $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a}) \geq 0 \qquad \forall\ \boldsymbol{a}, \boldsymbol{b} \in S$. Here, S is our space.

(ii) $d(\boldsymbol{a}, \boldsymbol{b}) = 0 \quad$ if and only if $\quad \boldsymbol{a} = \boldsymbol{b}$.

(iii) $d(\boldsymbol{a}, \boldsymbol{b}) \leq d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{c}, \boldsymbol{b})$.

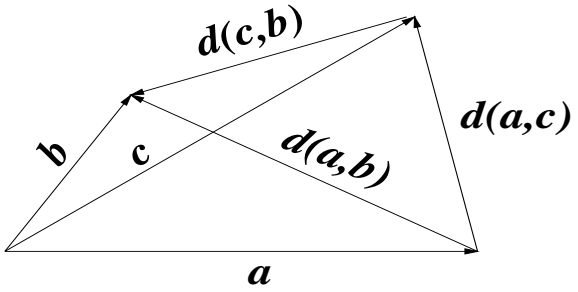


Fig. 7.7 Triangle inequality.

The third item is called *triangle inequality* illustrated in Fig. 7.7.

The norm is just the "distance" implied by the length of the vector $\boldsymbol{A} = \boldsymbol{a} - \boldsymbol{b}$. Then the norm of $\boldsymbol{A}$ we denote with two bars as

$$\ell = \|\boldsymbol{A}\|. \tag{7.38}$$

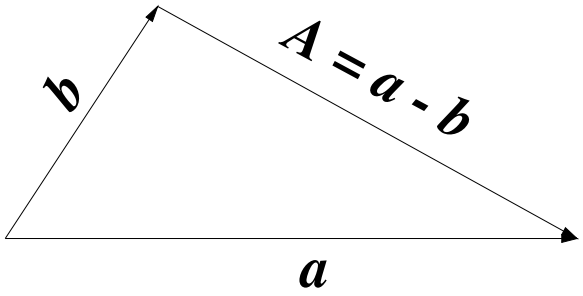


Fig. 7.8 Norm of a vector.

According to the above we have

(i) $\|\boldsymbol{A}\| > 0 \quad \text{unless} \quad \boldsymbol{A} = \boldsymbol{0}$.

(ii) $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.

We add the following as part of the definition

(iii) $\|k\,\boldsymbol{A}\| = k\,\|\boldsymbol{A}\|$.

We say that a norm "induces" a topology. In $\boldsymbol{R}^2$ the equation $\|\boldsymbol{a}\| = 1$ produces different shapes depending upon the geometry.

For ℓ_2 norm $\Longrightarrow$ $x^2 + y^2 = 1$ $\Longrightarrow$ circle;

For ℓ_1 norm $\Longrightarrow$ $|x| + |y| = 1$ $\Longrightarrow$ four straight lines with slopes $\pm$ 1;

For ℓ_∞ norm $\Longrightarrow$ $\max\{x, y\} = 1$ $\Longrightarrow$ in this case, $x = 1$ or $y = 1$.

The above geometrical meaning of norms is shown in Fig. 7.9.

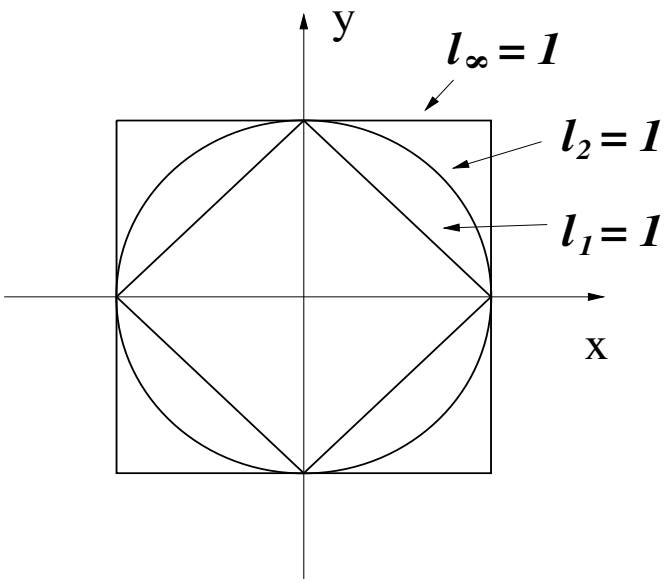


Fig. 7.9 Geometrical meaning of different norms.

7.4 Existence and Uniqueness Theorem

Having defined norms, we can now define *Lipschitz* vector functions. Suppose we have system of equations:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, \cdots, x_n; t) \\ \dot{x}_2 &= f_2(x_1, x_2, \cdots, x_n; t) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, x_2, \cdots, x_n; t) \end{aligned} \tag{7.39}$$

or in the vector form:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x};\, t). \tag{7.40}$$

We say f_i is continuous in $(\boldsymbol{x},\, t)$ if it is continuous in $x_1,\, x_2,\, \cdots,\, x_n$ and t. A function is said analytic in $x_1,\, x_2,\, \cdots,\, x_n;\, t$ if it can be expressed as a power series in the $n+1$ variables $x_1,\, x_2,\, \cdots,\, x_n$ and t.

A vector function $\boldsymbol{f}(\boldsymbol{x};\, t)$ is a *Lipschitz function* in $\boldsymbol{x}$ if there exists m such that

$$\|\boldsymbol{f}(\boldsymbol{x};\, t) - \boldsymbol{f}(\boldsymbol{y};\, t)\| \leq m\, \|\boldsymbol{x} - \boldsymbol{y}\| \tag{7.41}$$

where m is a finite constant, called a Lipschitz constant.

This turns out to be a less restrictive requirement than the existence of a derivative. We note that something like a mean value theorem makes function Lipschitz.

So far when we solve a D.E. we always assume or imply that solutions exist (even if we can't "get" them). However, the following simple example

$$\dot{\boldsymbol{x}}^2 + \frac{1}{2}\,\boldsymbol{x}^2 = -\,2$$

does not have any real solutions. So if for a general D.E. our solution techniques (e.g. perturbation approaches, direct integration methods, etc.) fail, how do we know whether there is actually anything else to look for? The following fundamental questions:

(1) When does the solution of a differential equation exist?

(2) Is the solution unique?

(3) Is the solution continuous on the initial condition $\boldsymbol{x}_0$?

are raised and they lead us to the most important results of the fundamental theory:

Theorem: (Existence and Uniqueness). Let $\boldsymbol{U} \subset \boldsymbol{R}^n$ be an open subset of Euclidean space; and let $\boldsymbol{f} : \boldsymbol{U} \longrightarrow \boldsymbol{R}^n$ be a C^1 map, and $\boldsymbol{x}_0 \in \boldsymbol{U}$. Then there exists some constant $a > 0$ and a unique solution

$$\boldsymbol{x} : (-\,a,\, a) \longrightarrow \boldsymbol{U}$$

of the D.E.

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{7.42}$$

satisfying the initial condition

$$\boldsymbol{x}(0) = x_0 \,.$$

Note: In fact all we need for $\boldsymbol{f}$ is that the function $\boldsymbol{f}(\boldsymbol{x})$ satisfies the Lipschitz condition:

$$\| \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}) \| \leq m \, \| \boldsymbol{x} - \boldsymbol{y} \| \qquad \forall \, \boldsymbol{x}, \, \boldsymbol{y} \in \boldsymbol{R}^n \tag{7.43}$$

where $\boldsymbol{f}$ is called a *Lipschitz function*. Here, the notation $\| \bullet \|$ denotes a **norm**. The Lipschitz condition turns out to be a less restrictive requirement than the existence of a derivative.

7.4.1. Existence Theorem. Let $\boldsymbol{U} \subset \boldsymbol{R}^n$ consisting of all $\boldsymbol{x}$ such that

$$\| \boldsymbol{x} - \boldsymbol{c} \| \leq a > 0 \tag{7.44}$$

where $\boldsymbol{c}$ is a constant vector defining a point in $\boldsymbol{U}$. Let $\boldsymbol{f}(\boldsymbol{x})$ be a continuous vector function in $\boldsymbol{U}$ satisfying the D.E. (7.42) for $0 \leq t \leq b$. We seek a solution such that $\boldsymbol{x}(0) = \boldsymbol{c}$ for $0 \leq t \leq b$. It should be noted that this "forward" problem also characterizes the problem in negative t by just changing the sign of $\boldsymbol{f}$.

In general, by integrating the D.E. (7.42) we have

$$\boldsymbol{x}(t) = \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{x}(\tau)) \, d\tau \tag{7.45}$$

on $0 \leq t \leq b$. It is helpful to think of the RHS of the equation as a transform, $T\{\boldsymbol{x}(t)\}$ that produces a new function on the LHS of the equation. The solution of the system of D.E.'s is a **fixed point** of this transform. Thus, for any $\boldsymbol{x}(t)$, let

$$T\{\boldsymbol{x}(t)\} = \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{x}(\tau)) \, d\tau = \boldsymbol{y}(t) \,, \tag{7.46}$$

then $\boldsymbol{x}(t)$ is a solution if $\boldsymbol{y}(t) = \boldsymbol{x}(t)$.

Now, since $\boldsymbol{f}(\boldsymbol{x})$ is continuous in $\boldsymbol{U}$, it is bounded in $\boldsymbol{U}$, thus

$$\| \boldsymbol{f}(\boldsymbol{x}) \| \leq K \quad \forall \, \boldsymbol{x} \in \boldsymbol{U}. \tag{7.47}$$

Therefore,

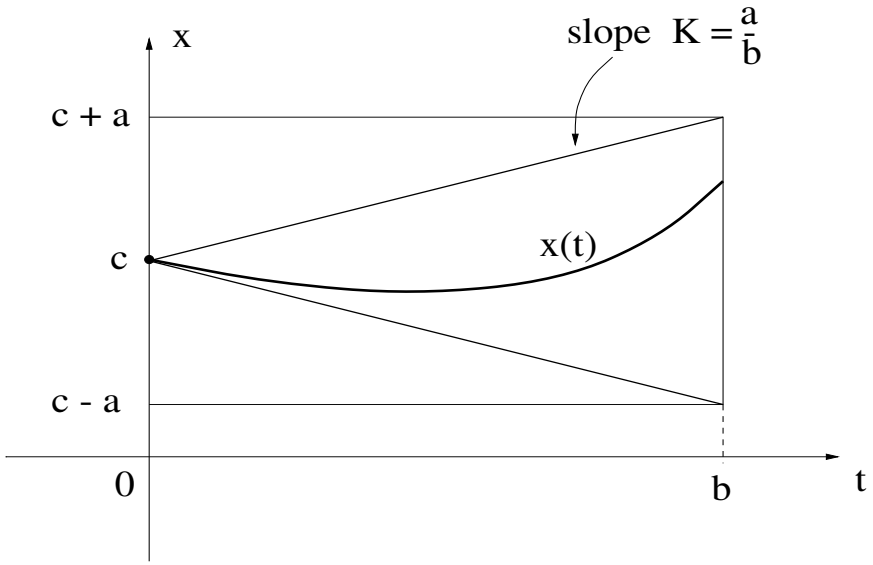


Fig. 7.10 Geometric explanation of the existence theorem.

$$\begin{aligned} \| \boldsymbol{y} - \boldsymbol{c} \| &= \| T\{\boldsymbol{x}\} - \boldsymbol{c} \| \\ &= \left\| \int_0^t \boldsymbol{f}(\boldsymbol{x}(\tau))\, d\tau \right\| \\ &\leq \int_0^t \| \boldsymbol{f}(\boldsymbol{x}(\tau)) \|\, d\tau \\ &\leq K\, t \end{aligned} \tag{7.48}$$

for any $\boldsymbol{x}$, $\boldsymbol{y} \in \boldsymbol{U}$.

If we choose b (recall that $0 \leq t \leq b$) such that

$$b = \frac{a}{K}, \tag{7.49}$$

then, the above inequality (7.48) implies that $\| \boldsymbol{y} - \boldsymbol{c} \| \leq a$ in addition to the requirement $\| \boldsymbol{x} - \boldsymbol{c} \| \leq a$ to define $\boldsymbol{U}$. That is, both input and output functions are members of $\boldsymbol{U}$ at least for $0 \leq t \leq \frac{a}{K}$. Consequently, $\boldsymbol{U}$ is mapped onto itself by T for $0 \leq t \leq \frac{a}{K}$ (see Fig. 2.17).

Fixed points of T clearly have this property that they are confined to $\boldsymbol{U}$ in that interval of t.

If T has no fixed points in $\boldsymbol{U}$, there is no solution;

If T has one fixed point in $\boldsymbol{U}$, there is one solution;

etc.

If we choose some function $\boldsymbol{x}^{(0)}(t)$ with appropriate behavior in $\boldsymbol{U}$ for $t \in [0,\, b]$ then T induces a sequence

$$\boldsymbol{x}^{(0)}(t)\,,\ \boldsymbol{x}^{(1)}(t)\,,\ \ldots\,,\ \boldsymbol{x}^{(n)}(t)\,,\ldots$$

which may converge to a solution (fixed point) of the original system. The following theorem guarantees the convergence of such a sequence. This method is called **successive approximation**.

Theorem: Let $\boldsymbol{x}^{(0)}(t)$ be a continuous function in $\boldsymbol{U}$ for $0 \le t \le b$. For $k = 0,\, 1,\, \ldots$ define

$$\boldsymbol{x}^{(k+1)}(t) = T\{\boldsymbol{x}^{(k)}(t)\} = \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(k)}(t))\, d\tau\,. \tag{7.50}$$

If $\boldsymbol{f}(\boldsymbol{x})$ is a Lipschitz function, then the sequence converges to a solution of the **initial value problem**.

PROOF: First note that

$$\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} = \boldsymbol{c} - \boldsymbol{x}^{(0)} + \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau))\, d\tau\,. \tag{7.51}$$

Thus, we have

$$\begin{aligned} \left\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right\| &\le \left\| \boldsymbol{c} - \boldsymbol{x}^{(0)} \right\| + \left\| \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau))\, d\tau \right\| \\ &\le \left\| \boldsymbol{c} - \boldsymbol{x}^{(0)} \right\| + \int_0^t \left\| \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau)) \right\| d\tau \\ &\le a + K\,t \\ &\le 2\,a \end{aligned} \tag{7.52}$$

where $K\,t \le a$ due to $b = \dfrac{a}{K}$, (i.e. $K \equiv \dfrac{a}{b}$) and $0 \le t \le b$.

Next, for $k = 1$,

$$\left\| \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \right\| = \left\| \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(1)}(\tau))\, d\tau - \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau))\, d\tau \right\|$$

$$
\begin{aligned}
&= \left\| \int_0^t \left[\boldsymbol{f}(\boldsymbol{x}^{(1)}(\tau)) - \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau)) \right] d\tau \right\| \\
&\leq \int_0^t \left\| \boldsymbol{f}(\boldsymbol{x}^{(1)}(\tau)) - \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau)) \right\| d\tau \\
&\leq \int_0^t m \left\| \boldsymbol{x}^{(1)}(\tau) - \boldsymbol{x}^{(0)}(\tau) \right\| d\tau \\
&\leq 2\,a\,(m\,t) \qquad (7.53)
\end{aligned}
$$

on $0 \leq t \leq b$. Finally, by induction, it follows that

$$
\left\| \boldsymbol{x}^{(k+1)}(t) - \boldsymbol{x}^{(k)}(t) \right\| \leq 2\,a\,\frac{(m\,t)^k}{k!} \leq 2\,a\,\frac{(m\,b)^k}{k!} \longrightarrow 0 \quad \text{as} \quad k \longrightarrow \infty.
$$

This shows that the sequence

$$
\boldsymbol{x}^{(0)}(t)\,,\ \boldsymbol{x}^{(1)}(t)\,,\ \ldots\,,\ \boldsymbol{x}^{(n)}(t)\,,\ldots
$$

uniformly (independence of time t) converges and the limit exists, hence, a solution exists. Thus, we have established under very general conditions that a solution must exist for the D.E. (7.42) **even** if we can't compute one.

7.4.2. Uniqueness Theorem. It is very important that only one solution passes through any particular point in phase space (uniqueness). This turns out to be a result of the Lipschitz condition. Uniqueness ultimately restricts very strongly what is possible and impossible in terms of solutions.

PROOF OF UNIQUENESS: Suppose, we have a solution $\boldsymbol{x}(t)$ (from the part of existence theorem), generated from the sequence $\boldsymbol{x}^{(0)}(t)\,,\ \boldsymbol{x}^{(1)}(t)\,,\ \ldots$, and assume that $\boldsymbol{y}(t)$ is another solution on the same interval $0 \leq t \leq b$. We want to prove that $\boldsymbol{y}(t) = \boldsymbol{x}(t)$.
From

$$
\begin{aligned}
\boldsymbol{x}^{(0)}(t) &= \boldsymbol{c} \\
\boldsymbol{x}^{(1)}(t) &= \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau))\,d\tau \\
&\vdots \\
\boldsymbol{x}^{(n)}(t) &= \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{x}^{(n-1)}(\tau))\,d\tau \\
\boldsymbol{y}(t) &= \boldsymbol{c} + \int_0^t \boldsymbol{f}(\boldsymbol{y}(\tau))\,d\tau\,, \qquad (7.54)
\end{aligned}
$$

we have

$$\begin{aligned}
\left\| \boldsymbol{x}^{(0)}(t) - \boldsymbol{y}(t) \right\| &= \left\| \int_0^t \boldsymbol{f}(\boldsymbol{y}(\tau))\, d\tau \right\| \leq K\, t \leq a\,, \\
\left\| \boldsymbol{x}^{(1)}(t) - \boldsymbol{y}(t) \right\| &= \left\| \int_0^t \left[\boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau)) - \boldsymbol{f}(\boldsymbol{y}(\tau)) \right] d\tau \right\| \\
&\leq \int_0^t \left\| \boldsymbol{f}(\boldsymbol{x}^{(0)}(\tau)) - \boldsymbol{f}(\boldsymbol{y}(\tau)) \right\| d\tau \\
&\leq \int_0^t m \left\| \boldsymbol{x}^{(0)}(\tau) - \boldsymbol{y}(\tau) \right\| d\tau \\
&\leq a\,(m\,t)\,. \qquad (7.55)
\end{aligned}$$

Again, by induction, one may obtain

$$\left\| \boldsymbol{x}^{(k)}(t) - \boldsymbol{y}(t) \right\| \leq a\,\frac{(m\,t)^k}{k!} \leq a\,\frac{(m\,b)^k}{k!} \;\longrightarrow\; 0 \quad \text{as} \quad k \;\longrightarrow\; \infty.$$

This implies that the sequence uniformly converges to $\boldsymbol{y}(t)$. But since $\boldsymbol{x}(t)$ is generated from this same sequence (i.e. the sequence also uniformly converges to $\boldsymbol{x}(t)$), therefore, $\boldsymbol{y}(t) = \boldsymbol{x}(t)$. This proves that the solution is unique.

Let us see how the "iteration scheme" used in the proof in this section can be applied to the following very simple example:

EXAMPLE 7.5.

$$\dot{x} = x.$$

We know that the solution, , $x(t)$, of the D.E. satisfying $x(0) = x_0$ is given by $x(t) = x_0\, e^t$).

Set $x^{(0)}(t) = x_0$, then by using formula (7.54) one obtains

$$\begin{aligned}
x^{(0)}(t) &= x_0 \\
x^{(1)}(t) &= x_0 + \int_0^t x^{(0)}(\tau)\, d\tau = x_0 + \int_0^t x_0\, d\tau = x_0 + x_0\, t = x_0\,(1+t) \\
x^{(2)}(t) &= x_0 + \int_0^t x^{(1)}(\tau)\, d\tau = x_0 + \int_0^t x_0\,(1+\tau)\, d\tau = x_0\,(1 + t + \frac{t^2}{2!}) \\
&\vdots \\
x^{(k)}(t) &= x_0 + \int_0^t x^{(k-1)}(\tau)\, d\tau = x_0\,(1 + t + \frac{t^2}{2!} + \cdots + \frac{t^k}{k!}) = x_0 \sum_{i=0}^{k} \frac{t^i}{i!}\,.
\end{aligned}$$

As $k \to \infty$, $x^{(k)}(t)$ converges to

$$x_0 \sum_{i=0}^{\infty} \frac{t^i}{i!} = x_0\, e^t = x(t),$$

which is, of course, the solution of the original D.E. $\dot{x} = x$.

Note: The Lipschitz condition definition (7.43), and the existence and uniqueness theory given above are applicable for autonomous systems only. However, they can be easily extended to non-autonomous systems. For example, the equations can be modified as

$$\| \boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y}) \| \leq m \, \| \boldsymbol{x} - \boldsymbol{y} \| \qquad \forall \ (t, \boldsymbol{x}), \ (t, \boldsymbol{y}) \in \boldsymbol{R} \times \boldsymbol{R}^n. \tag{7.56}$$

for non-autonomous systems. The proof of existence and uniqueness theorem for non-autonomous systems can be found from, for example, Hirsch and Small (1974, Chapter 15).

Also note that the theorem is **local**. Furthermore, in order for the theorem to be physically relevant, it needs to be complemented with the property that the solution $\boldsymbol{x}(t)$ depends continuously on the initial condition.

Remarks and Examples.

R1) An immediate implication is that cases such as shown in Fig. 2.18 are excluded by the theorem.

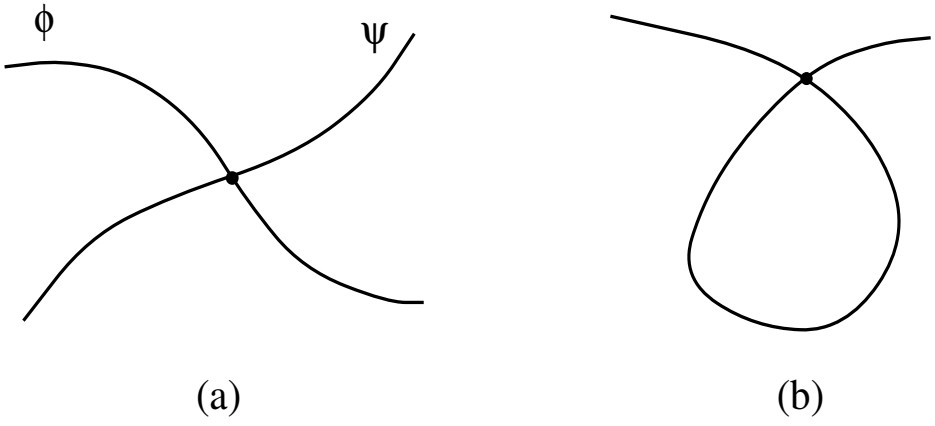


Fig. 7.11 Non-unique solutions: (a) two solutions cross; (b) a solution crosses itself.

Moreover, if a solution curve ϕ satisfies $\phi(t_1 + w) = \phi(t_1)$, then the curve must close up as shown in Fig. 2.19.

R2) Consider the IVP (Initial Value Problem)

$$\dot{x} = 1 + x^2 \ , \qquad x(0) = 0 \tag{7.57}$$

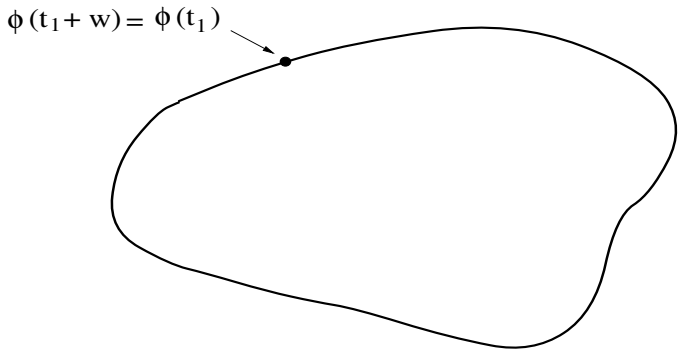


Fig. 7.12 A closed solution curve.

which may have solutions

$$\begin{aligned} x &= \tan t\ , \qquad -\frac{\pi}{2} \le t \le \frac{\pi}{2} \\ x &= \tan t\ , \qquad -\frac{1}{2} \le t \le \frac{1}{2} \end{aligned} \tag{7.58}$$

and, strictly speaking, they should be considered different. But this is artificial and inconvenient. To avoid this, we introduce the concept of **maximal** solutions, i.e. roughly speaking, solutions which hold over the largest possible interval. For the above example, the general solution is given by the function

$$x = \tan(t - c)\ , \qquad c = \text{constant}. \tag{7.59}$$

Such a function cannot be extended over an interval larger than

$$c - \frac{\pi}{2} < t < c + \frac{\pi}{2} \tag{7.60}$$

since $x(t) \longrightarrow \pm\infty$ as $c \pm \dfrac{\pi}{2}$.

R3) Note that **globality** does not require the solution to be defined for all t. Consider the following example

$$\dot{x} = x^2 \qquad \text{in } \boldsymbol{R} \qquad \text{with} \quad x(0) \equiv x_0 > 0 \tag{7.61}$$

The solution of this D.E. is $x = \dfrac{1}{\dfrac{1}{x_0} - t}$, defined for $-\infty < t < \dfrac{1}{x_0}$.

To end this chapter, we use the following example to demonstrate how to discuss the uniqueness of solutions of a D.E.

EXAMPLE 7.6. Consider the D.E

$$\frac{dx}{dt} = -K\sqrt{x}\,, \quad x \geq 0 \qquad (K > 0 \text{ is a constant}) \tag{7.62}$$

a) Find a value of x for which the D.E. does not have unique solutions. What are these solutions?

b) Why does this contradict the uniqueness theorem?

c) The D.E. represents the rate at which the depth of liquid in a bucket decreases due to a small leak. Explain how non-uniqueness pertains to this physical problem.

Solution:

(a) Consider the functions

$$x(t) \equiv 0 \qquad \text{and} \qquad x(t) = \frac{1}{4}K^2(t+C)^2 \qquad (t \leq -C) \tag{7.63}$$

where C is an arbitrary constant. When $x = 0$, for any given initial point $(t_0\,,\,0)$ in the $(t\,,\,x)$ plane, one can choose $C = -t_0$ so that there are two solutions given by (7.63) satisfying the D.E. and passing through the same point $(t_0\,,\,0)$. Thus, the solution $x = 0$ is not unique. The solutions of the D.E. (7.62) can be solved as follows:

$$\begin{aligned}
& \frac{dx}{dt} = -K\sqrt{x} \\
\Longrightarrow \quad & x^{-1/2}\,dx = -K\,dt \\
\Longrightarrow \quad & 2\,x^{1/2} = -K(t+C) \qquad (t \leq -C \ \text{ since } \ x \geq 0 \ \text{ and } \ K > 0) \\
\Longrightarrow \quad & x(t) = \frac{1}{4}K^2(t+C)^2 \qquad (t \leq -C).
\end{aligned} \tag{7.64}$$

The solutions given by (7.63) are shown in Fig. 7.13. Note that only half of the parabola are the family of the solutions $x(t) = \frac{1}{4}K^2(t+C)^2 \ \ (t \leq -C)$. It can be proved that the solution $x(t) \equiv 0$ is a **singular** solution, which is the **envelope** of the family of the solutions $\frac{1}{4}K^2(t+C)^2$.

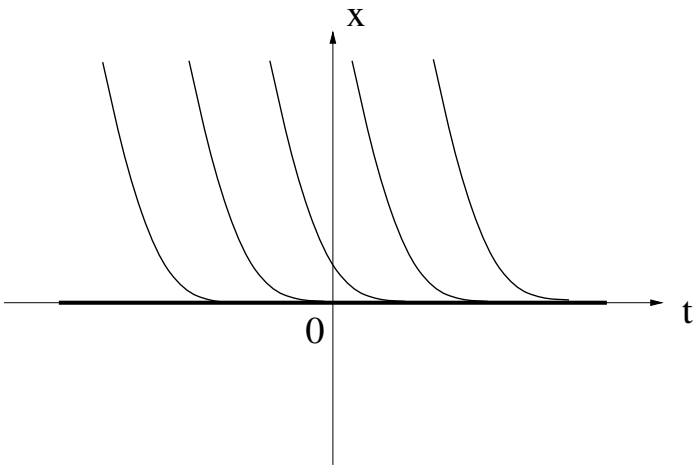


Fig. 7.13 Solutions of D.E (7.63).

(b) Because it does not satisfy the Lipschitz condition at $x = 0$. Here, $f(x) = -K\sqrt{x}$, so, for arbitrary x_1 and x_2 in the neighborhood of $x = 0$, we have

$$\begin{aligned} \| f(x_1) - f(x_2) \| &= | f(x_1) - f(x_2)| \\ &= |-K\sqrt{x_1} + K\sqrt{x_2}| \\ &= |K|\,|\sqrt{x_1} - \sqrt{x_2}| \\ &= |K| \left| \frac{\sqrt{x_1} - \sqrt{x_2}}{x_1 - x_2} \right| |x_1 - x_2| \\ &= |K| \left| \frac{1}{\sqrt{x_1} + \sqrt{x_2}} \right| |x_1 - x_2| \,. \end{aligned} \tag{7.65}$$

Want to the Lipschitz condition to be satisfied, there must exist a Lipschitz constant (finite) $m \geq 0$ such that

$$| f(x_1) - f(x_2)| \leq m\, |x_1 - x_2| \qquad \forall\, x_1\,,\ x_2 \geq 0. \tag{7.66}$$

It is clear that such an m does not exist because $\dfrac{|K|}{\sqrt{x_1} + \sqrt{x_2}} \to \infty$ when $x_1\,,\ x_2 \to 0$.

(c) One cannot determine when a formerly full, empty bucket began to leak. Empty bucket corresponds to the solution $x = 0$. Since this solution is not unique, so one cannot determine which branch represents the solution corresponding to the non-empty bucket.

7.5 Autonomous vs. Non-autonomous Systems

You may notice that we defined a Lipschitz function $\boldsymbol{f}(\boldsymbol{x},\, t)$ as

$$\|\boldsymbol{f}(\boldsymbol{x};\, t) - \boldsymbol{f}(\boldsymbol{y};\, t)\| \leq m\, \|\boldsymbol{x} - \boldsymbol{y}\|$$

where m is a Lipschitz constant. Note here function $\boldsymbol{f}(\boldsymbol{x},\, t)$ contains time t explicitly. But when we prove the existence and uniqueness theorem of DE we considered the system

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$$

where the vector function does not include time t explicitly. This type of systems is called *autonomous systems*. For example,

$$\dot{x} = x + 3\, x^3.$$

The dynamic behaviour of an autonomous system can be quite different from a *non-autonomous system* which is described in the general form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x},\, t).$$

For example,

$$\dot{x} = x + 3\, x^3 + \sin t\,.$$

The former is subject to the existence and uniqueness theorem, while the later is not generally necessarily unique even if $\boldsymbol{f}(\boldsymbol{x},\, const.)$ is Lipschitz in $\boldsymbol{x}$. This turns out to be very important for the question of chaos, as we shall see. It permits lower order systems than *three* to have chaos. For autonomous systems chaos is only possible for systems with *three or more dimensions*. It is possible to have an existence and uniqueness condition on a non-autonomous but generally only for a finite interval in time.

The reason for this becomes plain if one considers the following example: The DE

$$\ddot{x} + g(x,\, \dot{x}) = 0$$

is clearly an autonomous system. Here, x is a scalar variable. Corresponding to this system, let $x_1 = x$ and $x_2 = \dot{x}$, then we have a system described in a 2-D phase space:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\, g(x_1,\, x_2) \end{aligned} \qquad \text{or} \qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} x_2 \\ -\, g(x_1,\, x_2) \end{pmatrix}. \tag{7.67}$$

However, the following DE

$$\ddot{x} + g(x, \dot{x}) = h(t)$$

is non-autonomous. Again, let $x_1 = x$ and $x_2 = \dot{x}$, we have the system

$$\begin{array}{ll} \dot{x}_1 & = x_2 \\ \dot{x}_2 & = -g(x_1, x_2) + h(t) \end{array} \qquad \text{or} \qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, t) = \begin{pmatrix} x_2 \\ -g(x_1, x_2) + h(t) \end{pmatrix} \tag{7.68}$$

also described in a 2-D phase space. The following innocent addition to equation (7.68) has significant consequences. Define an additional variable

$$x_3 = t \qquad \text{then} \qquad \dot{x}_3 = 1,$$

so we can transform system (7.68) into a system

$$\begin{array}{ll} \dot{x}_1 & = x_2 \\ \dot{x}_2 & = -g(x_1, x_2) + h(x_3) \\ \dot{x}_3 & = 1 \end{array} \qquad \text{or} \qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, t) = \begin{pmatrix} x_2 \\ -g(x_1, x_2) + h(x_3) \\ 1 \end{pmatrix} \tag{7.69}$$

which is now a 3-D phase space and is an autonomous system. System (7.69) contains the original solution (x_1, x_2), but as a projection onto a subspace. Thus while (7.69) is fully subject to existence and uniqueness, system (7.68) may be not, because any unique solution path in the space of system (7.69) is a projection onto that of (7.68). This is illustrated in Fig. 7.14.

This point about the number of dimensions of the system is key here! We see that forcing increases the dimensionality of the system. It is not guaranteed that uniqueness will be violated in the original space, but it may be: that is, the original theorem are not violated if it is.

Remark: *Limitation of Existence and Uniqueness.* Two of the key requirements of the theorem are that solutions retain in a finite time interval and that the Lipschitz condition. It is not necessary that the later hold for all intervals in time t. For example, the DE

$$\dot{x} = 1 + x^2$$

has the general solution $x = \tan(t + C)$. This is guaranteed to have any finite space like $\boldsymbol{R}$ in a finite interval. Thus, globally the conditions need not hold, but locally they must. We say in this sense that this (Lipschitz) condition is local.

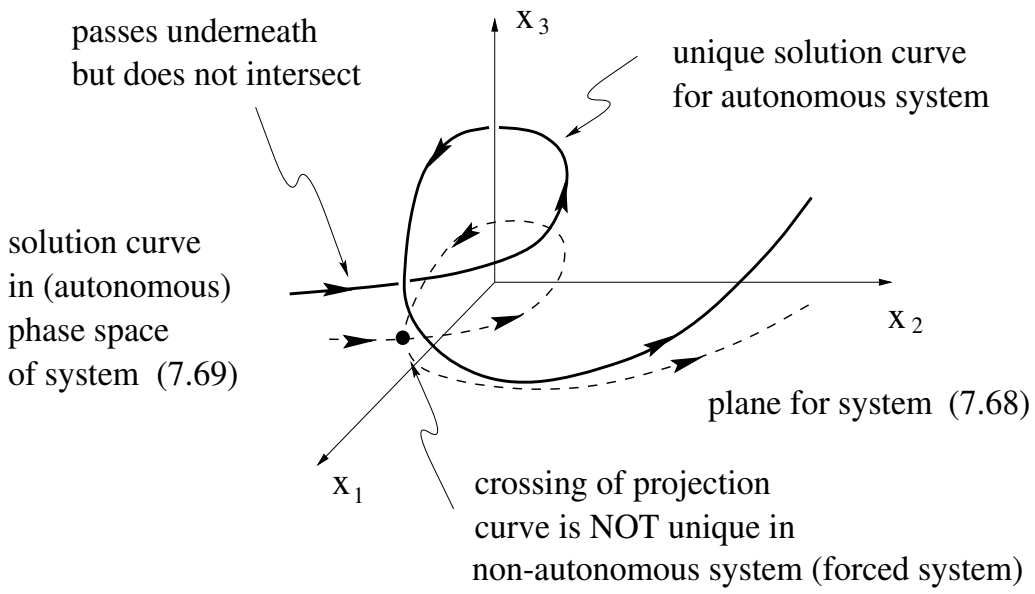


Fig. 7.14 Autonomous vs. non-autonomous systems.

This is not an issue for solutions that only trace out a finite region of space in $\boldsymbol{R}^n$ or solutions on "compact spaces" (e.g. torus).

Chapter 8

FIXED POINTS AND LINEARIZED STABILITY

We have already seen the importance of fixed points (singular points) in characterizing the nature of solutions of a nonlinear system such as the logistic (quadratic) map. The question we'll ask again about a DE is whether or not the fixed points of a DE are stable or unstable. In particular, how the trajectories (solutions curves) behave near the fixed points.

Since nearness is what we need, it follows that linearization about the fixed point can tell us a great deal about the behaviour in question. For the periodic case we'll linearize about a reference solution: both are conceptually straightforward.

Nonetheless, we'll find that the nonlinearities will introduce some cases anyway, making linear stability analysis meaningless. This has to do with the notion of structural stability.

8.1 Linearization

Linearization can be thought of in the sense of an approximation. Alternatively it can be thought of in terms "local" behavior of a system of equations.

It is not necessary to linearize a system about a fixed point: we can just as well discuss linearization about a reference function $\tilde{\boldsymbol{y}}(t)$. We suppose that there is some nearby trajectory $\boldsymbol{y}(t)$ such that

$$\boldsymbol{y}(t) = \tilde{\boldsymbol{y}}(t) + \boldsymbol{x}(t) \tag{8.1}$$

where $\|\boldsymbol{x}(t)\| < \epsilon$ for some small interval in time. Then, if

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}) \qquad \text{and thus} \qquad \dot{\tilde{\boldsymbol{y}}} = \boldsymbol{f}(\tilde{\boldsymbol{y}})$$

is the dynamical system being studied, we have $\boldsymbol{x} = \boldsymbol{y} - \tilde{\boldsymbol{y}}$ and by using Taylor series we obtain

$$\dot{\boldsymbol{y}} = \dot{\tilde{\boldsymbol{y}}} + \dot{\boldsymbol{x}} = \boldsymbol{f}(\tilde{\boldsymbol{y}}(t)) + D\boldsymbol{f}(\tilde{\boldsymbol{y}}(t))\,\boldsymbol{x} + O(\|\boldsymbol{x}\|^2) \tag{8.2}$$

where vector function $\boldsymbol{f}$ has n components f_i, $i = 1, 2, \cdots, n$; $D\boldsymbol{f}$ represents Jacobian matrix $\left[\dfrac{df_i}{dx_j}\right]_{\boldsymbol{y}=\tilde{\boldsymbol{y}}}$ The result of the product of this Jacobian matrix with $\boldsymbol{x}$ is the sum of all first derivatives of f_i, as expected. The notation $O(\|\boldsymbol{x}\|^2)$ denotes higher order terms starting from second-order.

We note since $\dot{\tilde{\boldsymbol{y}}} = \boldsymbol{f}(\tilde{\boldsymbol{y}})$, it follows from (8.2) that

$$\dot{\boldsymbol{x}} = D\boldsymbol{f}(\tilde{\boldsymbol{y}}(t))\,\boldsymbol{x} + O(\|\boldsymbol{x}\|^2). \tag{8.3}$$

This relation describes the condition of orbits (traced out trajectories) near $\tilde{\boldsymbol{y}}(t)$ due to perturbation. Our questions are addressed for orbits arbitrarily close to $\tilde{\boldsymbol{y}}(t)$, so we expect

$$\dot{\boldsymbol{x}} = D\boldsymbol{f}(\tilde{\boldsymbol{y}}(t))\,\boldsymbol{x} \tag{8.4}$$

to answer our questions. We refer to equation (8.4) as the **associated linear system** of the original nonlinear system.

EXAMPLE 8.1. Consider the (generalized) Duffing equation

$$\ddot{y} + \omega^2\, y + \mu\, y^3 = 0. \tag{8.5}$$

(Note: Duffing equation traditionally has slightly different signs $\ddot{y} - \omega^2\, y + |\mu|\, y^3 = 0$.) Introduce

$$y_1 = y\,, \qquad y_2 = \dot{y}$$

to transform (8.5) into the system

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\,\omega^2\, y_1 - \,\mu\, y_1^3 \end{pmatrix} \equiv \begin{pmatrix} f_1(y_1,\, y_2) \\ f_2(y_1,\, y_2) \end{pmatrix}. \tag{8.6}$$

Thus,

$$\frac{\partial f_i}{\partial y_j} = \begin{bmatrix} \dfrac{\partial f_1}{\partial y_1} & \dfrac{\partial f_1}{\partial y_2} \\ \dfrac{\partial f_2}{\partial y_1} & \dfrac{\partial f_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\,\omega^2 - 3\,\mu\, y_1^2 & 0 \end{bmatrix}. \tag{8.7}$$

So, the matrix associated with the linear system is given by (8.7), where y_1 is evaluated at a reference solution.

Now, in practice, fixed points are the typical reference solution, they occur when

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} y_2 \\ -\omega^2 y_1 - \mu y_1^3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hence, the fixed points are given by

$$(y_1,\, y_2) = (0,\, 0) \qquad \text{for any value of } \ \mu; \tag{8.8}$$

and

$$(y_1^{\pm},\, y_2) = \left(\pm \frac{\omega}{\sqrt{-\mu}},\, 0 \right) \qquad \text{for } \ \mu < 0 \ \text{ only.} \tag{8.9}$$

The linear system associated with the fixed point $(0,\, 0)$ is given by

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{8.10}$$

where $\boldsymbol{x}$ is the displacement orbit from the fixed point $(0,\, 0)$. Then eigenvalues of the matrix are key. The characteristic polynomial of the above equation is

$$\det \begin{bmatrix} \lambda & 1 \\ \omega^2 & \lambda \end{bmatrix} = 0 \qquad \Longrightarrow \qquad \lambda^2 + \omega^2 = 0 \tag{8.11}$$

which gives the eigenvalues $\lambda = \pm i\,\omega$, as we expected: The linearized system is, after all, none other than the simple harmonic oscillator!

We will discuss the meaning of these eigenvalues in the context of phase space below. In the mean time, we note that there are two questions we face here for an arbitrary $\tilde{\boldsymbol{y}}(t)$:

1) Is the case $\boldsymbol{x} = 0$ ($\tilde{\boldsymbol{y}}(t)$ in the original system) stable in the associated linearized system?

2) Does the stability of the $\boldsymbol{x} = 0$ of the linear problem imply that $\tilde{\boldsymbol{y}}(t)$ is stable?

8.2 Theory of Linear Systems

Suppose we are here concerned with the case where

$$\tilde{\boldsymbol{y}}(t) = \tilde{\boldsymbol{y}} = const. \tag{8.12}$$

That is, we expand the system about a fixed point. In this case

$$D\boldsymbol{f}(\tilde{\boldsymbol{y}}(t)) = D\boldsymbol{f}(\tilde{\boldsymbol{y}}) = A \tag{8.13}$$

where A is a constant matrix only. We thus have the associated linear system

$$\dot{\boldsymbol{x}} = A\,\boldsymbol{x}. \tag{8.14}$$

The solution of this linear system is easy in principle. After all, it is equivalent to nothing more than an nth order constant coefficient linear ODE, if $D\boldsymbol{f}$ is $n \times n$.

However, we need theory of linear systems to discuss linearization and, in fact, the linear system theory itself is an important part of dynamical systems. So, in following, we'll summarize some important results of linear systems, such as transformations, eigenvalues, eigenvectors and Jordan canonical form.

8.2.1. Solutions of Linear Differential Equations. Consider the linear DE (8.14)

$$\dot{\boldsymbol{x}} = A\,\boldsymbol{x}\,,$$

where A is a constant $n \times n$ matrix and $\boldsymbol{x}$ is an n-dimensional vector.

Define the following operation on the matrix A :

$$e^{At} = I + At + \frac{1}{2!}\,A^2t^2 + \frac{1}{3!}\,A^3t^3 + \cdots \tag{8.15}$$

where I is a unit matrix. For a square matrix A this series converges for every A and t. Thus, e^{At} is some matrix M. We note that

$$\frac{dM}{dt} = A\,M \quad (\text{where } M = e^{At}). \tag{8.16}$$

Thus $M\,\boldsymbol{k}$ is a good solution to try in (8.14) and indeed, substituting it into (8.14) yields

$$\text{LHS} = \frac{d}{dt}\,(M\,\boldsymbol{k}) = \frac{d}{dt}(M)\,\boldsymbol{k} = A\,M\,\boldsymbol{k} = \text{RHS}\,,$$

so, $M\,\boldsymbol{k}$ or $e^{At}\,\boldsymbol{k}$ is a solution. In fact, it turns out that for

$$\dot{\boldsymbol{x}} = A\,\boldsymbol{x} + \boldsymbol{g}(t)\,, \tag{8.17}$$

the general solution is

$$e^{At}\,\boldsymbol{k} + e^{At}\int e^{-At}\,\boldsymbol{g}(t)\,dt\,. \tag{8.18}$$

Thus, for the linear system (8.14), the general solution is given by

$$\boldsymbol{x} = e^{At}\,\boldsymbol{x}_0\,, \tag{8.19}$$

where $\boldsymbol{x}_0$ is the initial condition vector. This sounds pretty simple, but much has been burden on the exponential. In particular, how do you find M without doing an infinite number of $n \times n$ matrix multiplications?

What we really want is to see that $e^{At}\,\boldsymbol{x}_0$ corresponds to the simple forms that we are accustomed to seeing in nth-order linear constant coefficient D.Es, that is, exponentials.

The eigenvalue problem for A can be expressed as

$$(A - \lambda_i\, I)\,\boldsymbol{e}_i = \boldsymbol{0} \tag{8.20}$$

where $\boldsymbol{e}_i$ is an eigenvector corresponding to the distinct eigenvalues. We form the matrix

$$T = (\boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \cdots \;\; \boldsymbol{e}_n) \tag{8.21}$$

composed of column eigenvectors. Then, it is clear that

$$\begin{aligned} AT &= (A\,\boldsymbol{e}_1 \; A\,\boldsymbol{e}_2 \; \cdots \; A\,\boldsymbol{e}_n) \\ &= (\lambda_1\,\boldsymbol{e}_1 \; \lambda_2\,\boldsymbol{e}_2 \; \cdots \; \lambda_n\,\boldsymbol{e}_n) \\ &= T \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \end{aligned} \tag{8.22}$$

Suppose that T has an inverse T^{-1}, then we have

$$T^{-1}AT = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \equiv \Lambda\,. \tag{8.23}$$

It follows that $A = T\Lambda T^{-1}$. Now, the general solution $\boldsymbol{x}$ of (8.14) becomes

$$\boldsymbol{x} = e^{At}\,\boldsymbol{x}_0 = e^{T\Lambda T^{-1}t}\,\boldsymbol{x}_0\,. \tag{8.24}$$

Since $e^{T\Lambda T^{-1}t} = I + T\Lambda T^{-1}\,t + \frac{1}{2!}\left(T\Lambda T^{-1}\right)^2 t^2 + \cdots$ and we note that

$$\begin{aligned}
\left(T\Lambda T^{-1}\right)^2 &= \left(T\Lambda T^{-1}\right)\left(T\Lambda T^{-1}\right) \\
&= T\Lambda\left(T^{-1}T\right)\Lambda T^{-1} \\
&= T\Lambda^2 T^{-1} \\
&\vdots \\
\left(T\Lambda T^{-1}\right)^m &= T\Lambda^m T^{-1}, \qquad (8.25)
\end{aligned}$$

and with

$$\Lambda^m = \begin{bmatrix} \lambda_1^m & 0 & \cdots & 0 \\ 0 & \lambda_2^m & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n^m \end{bmatrix}, \qquad (8.26)$$

we have

$$\begin{aligned}
e^{T\Lambda T^{-1}t} &= T\begin{bmatrix} 1+\lambda_1 t+\frac{1}{2!}\lambda_1^2 t^2+\cdots & 0 & \cdots & 0 \\ 0 & 1+\lambda_2 t+\frac{1}{2!}\lambda_2^2 t^2+\cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1+\lambda_n t+\frac{1}{2!}\lambda_n^2 t^2+\cdots \end{bmatrix} T^{-1} \\
&= T\begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} T^{-1} \\
&= \left(\, \boldsymbol{e}_1 e^{\lambda_1 t} \;\; \boldsymbol{e}_2 e^{\lambda_2 t} \;\; \cdots \;\; \boldsymbol{e}_n e^{\lambda_n t} \,\right) T^{-1}. \qquad (8.27)
\end{aligned}$$

Finally,

$$\boldsymbol{x} = \left(\, \boldsymbol{e}_1\, e^{\lambda_1 t} \;\; \boldsymbol{e}_2\, e^{\lambda_2 t} \;\; \cdots \;\; \boldsymbol{e}_n\, e^{\lambda_n t} \,\right) T^{-1}\, \boldsymbol{x}_0 \qquad (8.28)$$

where $\boldsymbol{e}_i$'s are vectors and T^{-1} is constant matrix, and $\boldsymbol{x}_0$ is a constant vector. Thus we see that time growth of $\boldsymbol{x}(t)$ as a vector is entirely dependent upon $\lambda_1\,,\ \lambda_2\,, \cdots,\ \lambda_n$. In particular, it is **dependent upon the signs of the real parts of** $\lambda_1\,,\ \lambda_2\,, \cdots,\ \lambda_n$.

In the case of repeated roots, the situation is not so simple. We need to find so called *generalized eigenvectors* to construct the transformation matrix, leading us to the theory of **transformations** and **Jordan canonical forms**.

8.2.2. Distinct Eigenvectors. First let's discuss when A has distinct eigenvalues $\lambda_1\,,\ \lambda_2\,,\ \cdots\ \lambda_n\,.$

Case I: $A \in C^{n\times n}$, where $C^{n\times n}$ denotes a complex product space. For this case there exists a transformation matrix T such that

$$T^{-1}A\,T = \bar{A} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \tag{8.29}$$

Here, λ_i are complex eigenvalues which are not necessarily complex conjugates, and T is given by

$$T = (\,\boldsymbol{e}_1\ \ \boldsymbol{e}_2\ \ \cdots\ \ \boldsymbol{e}_n\,)$$

where $\boldsymbol{e}_i$ are solved from the eigenvector equations:

$$(A - \lambda_i\, I)\,\boldsymbol{e}_i = \mathbf{0}\,,$$

for $\boldsymbol{e}_i$ an eigenvector associated with the eigenvalue $\lambda_i\,.$

EXAMPLE 8.2. Given $A = \begin{bmatrix} -3 & -1 \\ 1-3\,i & -2\,i \end{bmatrix}$, where $i = \sqrt{-1}$, find the Jordan canonical form.

Solution: The characteristic polynomial $G(\lambda)$ is

$$\begin{aligned} G(\lambda) &= \det(A - \lambda\, I) \\ &= \det \begin{bmatrix} -\lambda - 3 & -1 \\ 1-3\,i & -\lambda - 2\,i \end{bmatrix} \\ &= \lambda^2 + (3+2\,i)\,\lambda + 1 + 3\,i \\ &= \lambda^2 + (3+2\,i)\,\lambda + (1+i)\,(2+i) \\ &= [\lambda + (1+i)]\ [\lambda + (2+i)]\ . \end{aligned} \tag{8.30}$$

Thus, the eigenvalues are $\lambda_1 = -1-i$ and $\lambda_2 = -2-i$. Substituting λ_1 into the eigenvector equation we obtain

$$\begin{bmatrix} -2+i & -1 \\ 1-3\,i & 1-i \end{bmatrix} \boldsymbol{e}_1 = \mathbf{0} \tag{8.31}$$

which, in turn, yields one independent equation (the determinant of the matrix is zero), as expected. Solving one of the algebraic equations yields a nonzero vector

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ -2+i \end{pmatrix} \qquad \text{(solution is not unique).} \tag{8.32}$$

Similarly, one can find $\boldsymbol{e}_2 = \begin{pmatrix} 1 \\ -1+i \end{pmatrix}$. Thus, the transformation matrix $T = (\boldsymbol{e}_1 \;\; \boldsymbol{e}_2)$ $= \begin{bmatrix} 1 & 1 \\ -2+i & -1+i \end{bmatrix}$. Then

$$T^{-1} = \frac{1}{-1+i-(-2+i)} \begin{bmatrix} -1+i & -1 \\ -(-2+i) & 1 \end{bmatrix} = \begin{bmatrix} -1+i & -1 \\ 2-i & 1 \end{bmatrix},$$

and therefore,

$$\begin{aligned} \bar{A} &= T^{-1} A T \\ &= \begin{bmatrix} -1+i & -1 \\ 2-i & 1 \end{bmatrix} \begin{bmatrix} -3 & -1 \\ 1-3\,i & -2\,i \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -2+i & -1+i \end{bmatrix} \\ &= \begin{bmatrix} 2 & 1+i \\ -5 & -2-i \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -2+i & -1+i \end{bmatrix} \\ &= \begin{bmatrix} -1-i & 0 \\ 0 & -2-i \end{bmatrix} \end{aligned} \tag{8.33}$$

which is in the Jordan canonical form.

Case II: $A \in R^{n\times n}$, where $R^{n\times n}$ represents a real product space. If $\lambda_i = a_i + i\,b_i$ is an eigenvector of A, then $\lambda_i^* = a_i - i\,b_i$ is also an eigenvector of A. Here, the superscript $*$ indicates the complex conjugate. Suppose $\boldsymbol{e}_i$ is the eigenvector associated with λ_i, then $\boldsymbol{e}_i^*$ is the eigenvector associated with λ_i^*. Assume that the eigenvalues λ_i are not repeated.

(i) Complex analysis: A transformation matrix $T_C \in C^{n\times n}$ can be found such that

$$T_C^{-1} A\, T_C = \bar{A}_C = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_1^* & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_k & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_k^* & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{2k+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{8.34}$$

where the subscript C denotes a complex transformation, and λ_1 , λ_1^* , $\cdots$ λ_k , λ_k^* are complex conjugate eigenvalues, while λ_{2k+1} , $\cdots$ λ_n are real eigenvalues, $k \leq \left[\frac{n}{2}\right]$. The transformation matrix T is given by

$$T_C = (\, \boldsymbol{e}_1 \;\; \boldsymbol{e}_1^* \;\; \cdots \;\; \boldsymbol{e}_k \;\; \boldsymbol{e}_k^* \,\vdots\, \boldsymbol{e}_{2k+1} \;\; \cdots \;\; \boldsymbol{e}_n \,), \tag{8.35}$$

where the first $2k$ eigenvectors are complex and the remaining are real.

EXAMPLE 8.3. Given $A = \begin{bmatrix} -1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & -1 & -1 \end{bmatrix}$, find the Jordan canonical form (in complex space).

Solution: The characteristic polynomial $G(\lambda)$ is

$$\begin{aligned} G(\lambda) &= \det(A - \lambda I) \\ &= \det \begin{bmatrix} -\lambda - 1 & 2 & 1 \\ 0 & -\lambda - 1 & 1 \\ 0 & -1 & -\lambda - 1 \end{bmatrix} \\ &= -(\lambda + 1)(\lambda^2 + 2\lambda + 2). \end{aligned} \tag{8.36}$$

Thus, the eigenvalues are $\lambda_1 = -1 + i$, $\lambda_2 = -1 - i$ and $\lambda_3 = -1$. The similar procedure used in example 1 can be applied here to find the eigenvectors:

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 - 2i \\ 1 \\ i \end{pmatrix}, \quad \boldsymbol{e}_2 = \begin{pmatrix} 1 + 2i \\ 1 \\ -i \end{pmatrix} \quad \text{and} \quad \boldsymbol{e}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

So, the transformation matrix is $T_C = (\boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \boldsymbol{e}_3) = \begin{bmatrix} 1 - 2i & 1 + 2i & 1 \\ 1 & 1 & 0 \\ i & -i & 0 \end{bmatrix}$.

Then $T_C^{-1} = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2}i \\ 0 & \frac{1}{2} & \frac{1}{2}i \\ 1 & -1 & 2 \end{bmatrix}$, and therefore,

$$\bar{A}_C = T_C^{-1} A T_C = \begin{bmatrix} -1 + i & 0 & 0 \\ 0 & -1 - i & 0 \\ 0 & 0 & -1 \end{bmatrix} \text{ which is in the canonical form.}$$

(ii) Real analysis: A transformation matrix $T_R \in R^{n\times n}$ can be found such that

$$T_R^{-1} A\, T_R = \bar{A}_R = \begin{bmatrix} a_1 & b_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ -b_1 & a_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_k & b_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -b_k & a_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{2k+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{8.37}$$

where the subscript R indicates a real transformation, and a_i and b_i are real, and the transformation matrix T_R is given by

$$T_R = (\,\mathrm{Re}(\boldsymbol{e}_1)\;\; \mathrm{Im}(\boldsymbol{e}_1)\;\; \cdots\;\; \mathrm{Re}(\boldsymbol{e}_k)\;\; \mathrm{Im}(\boldsymbol{e}_k)\;\; \boldsymbol{e}_{2k+1}\;\; \cdots\;\; \boldsymbol{e}_n\,)\,. \tag{8.38}$$

PROOF: From the results of complex analysis, we have

$$T_C^{-1} A\, T_C = \bar{A}_C \tag{8.39}$$

which can be written as

$$A\, T_C = T_C \bar{A}_C \tag{8.40}$$

or

$$A(\boldsymbol{e}_1\;\; \boldsymbol{e}_1^*\;\; \cdots\;\; \boldsymbol{e}_k\;\; \boldsymbol{e}_k^*\;\; \vdots\;\; \boldsymbol{e}_{2k+1}\;\; \cdots\;\; \boldsymbol{e}_n)$$

$$= (\boldsymbol{e}_1\;\; \boldsymbol{e}_1^*\;\; \cdots\;\; \boldsymbol{e}_k\;\; \boldsymbol{e}_k^*\;\; \vdots\;\; \boldsymbol{e}_{2k+1}\;\; \cdots\;\; \boldsymbol{e}_n) \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_1^* & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_k & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_k^* & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{2k+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

which can be rewritten as

$$(A\boldsymbol{e}_1\;\; A\boldsymbol{e}_1^*\;\; \cdots\;\; A\boldsymbol{e}_k\;\; A\boldsymbol{e}_k^*\;\; \vdots\;\; A\boldsymbol{e}_{2k+1}\;\; \cdots\;\; A\boldsymbol{e}_n)$$

$$= (\lambda_1\boldsymbol{e}_1\;\; \lambda_1^*\boldsymbol{e}_1^*\;\; \cdots\;\; \lambda_k\boldsymbol{e}_k\;\; \lambda_k^*\boldsymbol{e}_k^*\;\; \vdots\;\; \boldsymbol{e}_{2k+1}\;\; \cdots\;\; \lambda_n\boldsymbol{e}_n)\,.$$

Now, let's consider the block

$$\left(A\boldsymbol{e}_j\;\; A\boldsymbol{e}_j^*\right) = \left(\lambda_j\boldsymbol{e}_j\;\; \lambda_j^*\boldsymbol{e}_j^*\right), \quad j = 1,\, 2,\, \cdots,\, k. \tag{8.41}$$

Let

$$\begin{aligned} \lambda_j &= a_j + i\, b_j \\ \lambda_j^* &= a_j - i\, b_j \qquad (j = 1,\, 2,\, \cdots,\, k) \end{aligned} \tag{8.42}$$

and

$$\begin{aligned} \boldsymbol{e}_j &= \boldsymbol{R}_j + i\, \boldsymbol{I}_j \\ \boldsymbol{e}_j^* &= \boldsymbol{R}_j - i\, \boldsymbol{I}_j \qquad (j = 1,\, 2,\, \cdots,\, k) \end{aligned} \tag{8.43}$$

where $\boldsymbol{R}_j$ and $\boldsymbol{I}_j$ represent real and imaginary parts, respectively. Then, the above block under consideration becomes

$$\begin{aligned} & (A(\boldsymbol{R}_j + i\, \boldsymbol{I}_j)\,,\ A(\boldsymbol{R}_j - i\ \boldsymbol{I}_j)) \\ = \ & ((a_j + i\ b_j)\,(\boldsymbol{R}_j + i\, \boldsymbol{I}_j)\,,\ (a_j - i\ b_j)\,(\boldsymbol{R}_j - i\, \boldsymbol{I}_j)) \\ = \ & ((a_j \boldsymbol{R}_j - b_j \boldsymbol{I}_j) + i\,(b_j \boldsymbol{R}_j + a_j \boldsymbol{I}_j)\,,\ (a_j \boldsymbol{R}_j - b_j \boldsymbol{I}_j) - i\,(b_j \boldsymbol{R}_j + a_j \boldsymbol{I}_j))\ . \end{aligned}$$

Balancing the real and imaginary parts (note that A is a real matrix) yields

$$(A\boldsymbol{R}_j\,,\ A\boldsymbol{R}_j) = ((a_j \boldsymbol{R}_j - b_j \boldsymbol{I}_j)\,,\ (a_j \boldsymbol{R}_j - b_j \boldsymbol{I}_j)) \tag{8.44}$$

and

$$(A\boldsymbol{I}_j\,,\ -A\boldsymbol{I}_j) = ((b_j \boldsymbol{R}_j + a_j \boldsymbol{I}_j)\,,\ -(b_j \boldsymbol{R}_j + a_j \boldsymbol{I}_j)) \tag{8.45}$$

which turns out to give

$$A\boldsymbol{R}_j = a_j \boldsymbol{R}_j - b_j \boldsymbol{I}_j \qquad \text{and} \qquad A\boldsymbol{I}_j = b_j \boldsymbol{R}_j + a_j \boldsymbol{I}_j\,. \tag{8.46}$$

Combining these two equations gives

$$A\,(\boldsymbol{R}_j \ \ \boldsymbol{I}_j) = (\boldsymbol{R}_j \ \ \boldsymbol{I}_j) \begin{bmatrix} a_j & b_j \\ -b_j & a_j \end{bmatrix} \qquad (j = 1,\, 2,\, \cdots,\, k). \tag{8.47}$$

Finally, by combining all such blocks together, one obtains

$$A(\boldsymbol{R}_1\ \boldsymbol{I}_1\ \cdots\ \boldsymbol{R}_k\ \boldsymbol{I}_k \,\vdots\, \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n)$$

$$= (\boldsymbol{R}_1\ \boldsymbol{I}_1\ \cdots\ \boldsymbol{R}_k\ \boldsymbol{I}_k \vdots \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n) \begin{bmatrix} a_1 & b_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ -b_1 & a_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_k & b_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -b_k & a_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{2k+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

which indicates a transformation matrix T_R, given by

$$T_R = (\boldsymbol{R}_1\ \boldsymbol{I}_1\ \cdots\ \boldsymbol{R}_k\ \boldsymbol{I}_k \vdots \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n) \tag{8.48}$$

such that

$$T_R^{-1} A T_R = \bar{A}_R = \begin{bmatrix} a_1 & b_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ -b_1 & a_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_k & b_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -b_k & a_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda_{2k+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{8.49}$$

which is in the real Jordan canonical form.

The relation between the transformation matrices T_C and T_R can be found as follows:

$$\begin{aligned} T_C &= (\,\boldsymbol{e}_1\ \boldsymbol{e}_1^*\ \cdots\ \boldsymbol{e}_k\ \boldsymbol{e}_k^* \vdots \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n\,) \\ &= (\,\boldsymbol{R}_1 + i\,\boldsymbol{I}_1\ \ \boldsymbol{R}_1 - i\,\boldsymbol{I}_1\ \ \boldsymbol{R}_2 + i\,\boldsymbol{I}_2\ \ \boldsymbol{R}_2 - i\,\boldsymbol{I}_2\ \cdots\ \boldsymbol{R}_k + i\,\boldsymbol{I}_k\ \ \boldsymbol{R}_k - i\,\boldsymbol{I}_k \vdots \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n\,) \\ &= (\,\boldsymbol{R}_1\ \boldsymbol{I}_1\ \cdots\ \boldsymbol{R}_k\ \boldsymbol{I}_k \vdots \boldsymbol{e}_{2k+1}\ \cdots\ \boldsymbol{e}_n\,) \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ i & -i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & i & -i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{aligned}$$

$$= T_R \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ i & -i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & i & -i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{8.50}$$

or

$$T_R = T_C \begin{bmatrix} \frac{1}{2} & -\frac{1}{2}i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{2} & \frac{1}{2}i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{2} & -\frac{1}{2}i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \frac{1}{2} & \frac{1}{2}i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{8.51}$$

Now, suppose a general system is

$$\dot{\boldsymbol{z}} = A\,\boldsymbol{z} + \boldsymbol{f}(\boldsymbol{z}), \tag{8.52}$$

where $\boldsymbol{f}(\boldsymbol{z})$ is a general nonlinear vector functions. By choosing the complex transformation

$$\boldsymbol{z} = T_C\,\boldsymbol{x}_C, \tag{8.53}$$

one can find the new system

$$\dot{\boldsymbol{x}}_C = T_C^{-1} A T_C\,\boldsymbol{x}_C + T_C^{-1}\,\boldsymbol{f}(T_C\,\boldsymbol{x}_C). \tag{8.54}$$

If the real transformation

$$\boldsymbol{z} = T_R\,\boldsymbol{x}_R, \tag{8.55}$$

then (8.52) becomes

$$\dot{\boldsymbol{x}}_R = T_R^{-1} A T_R\,\boldsymbol{x}_R + T_R^{-1}\,\boldsymbol{f}(T_R\,\boldsymbol{x}_R). \tag{8.56}$$

Comparing equation (8.53) with (8.55) yields

$$T_C\,\boldsymbol{x}_C = T_R\,\boldsymbol{x}_R \qquad \Longrightarrow \qquad \boldsymbol{x}_C = T_C^{-1}\,T_R\,\boldsymbol{x}_R. \tag{8.57}$$

Substituting $\boldsymbol{x}_c = T_C^{-1} T_R \boldsymbol{x}_R$ into (8.54) results in (8.56). Let

$$P = T_C^{-1} T_R = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2}i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{2} & \frac{1}{2}i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{2} & -\frac{1}{2}i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \frac{1}{2} & \frac{1}{2}i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{8.58}$$

and then

$$P^{-1} = \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ i & -i & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & i & -i & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{8.59}$$

EXAMPLE 8.3 (continued). From complex analysis, we can obtain the real transformation matrix $T_R = (\mathrm{Re}(\boldsymbol{e}_1) \;\; \mathrm{Im}(\boldsymbol{e}_1) \;\; \boldsymbol{e}_3) = \begin{bmatrix} 1 & -2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.

Then $T_R^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 2 \end{bmatrix}$, and therefore,

$$\begin{aligned} \bar{A}_R &= T_R^{-1} A T_R \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} -1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & -1 & 1 \\ 0 & -1 & -1 \\ -1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{8.60}$$

which is in the real Jordan canonical form.

Case III: $A \in R^{n\times n}$ and is symmetric, that is $A = A^t$, where the superscript t denotes the transpose. For this case, there exists an orthogonal transformation matrix T such that $T^t A T = \bar{A}$ which is in the Jordan canonical form. Here $T^t = T^{-1}$, i.e. $T\,T^t = I$.

In the above, we discussed when A has distinct eigenvalues. Now suppose that the roots of the characteristic polynomial may be repeated. In this case, we may or may not find n independent eigenvectors which are needed to construct the transformation matrix.

8.2.3. Repeated Eigenvalues. If the eigenvalues of A are not all distinct, then it is not always possible to find a diagonal matrix representation (in complex space).

Case I: A is diagonalizable. Consider the following example.

EXAMPLE 8.4. $A = \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$.

Solution: The eigenvalues are $\lambda_1 = \lambda_2 = -1$, and $\lambda_3 = -2$. $\lambda = -1$ is a repeated eigenvalue. The eigenvectors associated with the eigenvalue λ_1 and λ_2 can be solved from the eigenvector equation $(A - \lambda_1 I)\,\boldsymbol{e} = \mathbf{0}$, i.e.

$$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \boldsymbol{e} = \mathbf{0}. \tag{8.61}$$

Note that the **rank** of the matrix is 1, so there is only one independent equation (three linear algebraic equations in total). Thus, there exist two linearly independent solutions, which are the eigenvectors we need:

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{8.62}$$

The eigenvector associated with λ_3 can be found as $\boldsymbol{e}_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$. Thus, the transformation

matrix $T = (\boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \boldsymbol{e}_3) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Then $T^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and therefore,

$$\begin{aligned} \bar{A} &= T^{-1}AT \\ &= \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \end{aligned} \tag{8.63}$$

which is in the Jordan canonical form. In this example, A has repeated eigenvalues but it is still diagonalizable. However, this is not always true. Let's look at another example.

Case II: A is not diagonalizable.

EXAMPLE 8.5. $A = \begin{bmatrix} -1 & 1 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & -2 \end{bmatrix}$.

Solution: The eigenvalues are $\lambda_1 = \lambda_2 = -1$, and $\lambda_3 = -2$. An eigenvector associated with λ_3 can be found as $\boldsymbol{e}_3 = \begin{pmatrix} 1 \\ -3 \\ 1 \end{pmatrix}$. The eigenvector equation for the repeated eigenvalue -1 is now given by

$$\begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix} \boldsymbol{e} = \boldsymbol{0}. \tag{8.64}$$

Now, the rank of the matrix is 2, so there are two independent equations. In other words, the **null** space $(A - \lambda_1 I)$ has dimension 1. Thus, there exists only one independent solution. Hence, in this case, we can only find one independent eigenvector, say, $\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ associated with the repeated eigenvalue. So we cannot construct the transformation matrix by using the independent eigenvectors because the number of column vectors needed is less than n. (In this example, n is 3, but we only obtain 2 independent eigenvectors.) In order

to construct the transformation matrix, we need to define **generalized eigenvectors**.

A vector $\boldsymbol{e}$ is said to be a generalized eigenvector of rank k of an $n \times n$ matrix A associated with eigenvalue λ_j if and only if

$$\begin{aligned} (A - \lambda_j I)^k \boldsymbol{e}_j &= \mathbf{0} \\ (A - \lambda_j I)^{k-1} \boldsymbol{e}_j &\neq \mathbf{0}. \end{aligned} \tag{8.65}$$

Note that if $k = 1$, it reduces to $(A - \lambda_j I)\boldsymbol{e}_j = \mathbf{0}$ and $\boldsymbol{e}_j \neq \mathbf{0}$, which is the definition of an eigenvector. Hence, the term *generalized eigenvector* is well-justified.

Let $\boldsymbol{e}_j$ be a generalized eigenvector of rank k associated with the eigenvalue λ_j. Define

$$\begin{aligned} \boldsymbol{e}_j^{(k)} &= \boldsymbol{e}_j \\ \boldsymbol{e}_j^{(k-1)} &= (A - \lambda_j I)\, \boldsymbol{e}_j = (A - \lambda_j I)\, \boldsymbol{e}_j^{(k)} \\ \boldsymbol{e}_j^{(k-2)} &= (A - \lambda_j I)^2\, \boldsymbol{e}_j = (A - \lambda_j I)\, \boldsymbol{e}_j^{(k-1)} \\ &\vdots \\ \boldsymbol{e}_j^{(2)} &= (A - \lambda_j I)^{k-2}\, \boldsymbol{e}_j = (A - \lambda_j I)\, \boldsymbol{e}_j^{(3)} \\ \boldsymbol{e}_j^{(1)} &= (A - \lambda_j I)^{k-1}\, \boldsymbol{e}_j = (A - \lambda_j I)\, \boldsymbol{e}_j^{(2)}. \end{aligned} \tag{8.66}$$

Then, for each i, $1 \leq i \leq k$, $\boldsymbol{e}_j^{(i)}$ is a generalized eigenvector of rank i. For example,

$$(A - \lambda_j I)^{k-2}\, \boldsymbol{e}_j^{(k-2)} = (A - \lambda_j I)^{k-2}\, (A - \lambda_j I)^2\, \boldsymbol{e}_j = (A - \lambda_j I)^k\, \boldsymbol{e}_j = \mathbf{0},$$

however,

$$(A - \lambda_j I)^{k-3}\, \boldsymbol{e}_j^{(k-2)} = (A - \lambda_j I)^{k-1}\, \boldsymbol{e}_j \neq \mathbf{0}.$$

Hence, $\boldsymbol{e}_j^{(k-2)}$ is a generalized eigenvector of rank of $k-2$ of A associated with the eigenvalue λ_j. We call $\left\{\boldsymbol{e}_j^{(1)},\ \boldsymbol{e}_j^{(2)},\ \cdots\ \boldsymbol{e}_j^{(k)}\right\}$ a chain of generalized eigenvectors.

Remarks. (i) The set of generalized eigenvectors $\left\{\boldsymbol{e}_j^{(1)},\ \boldsymbol{e}_j^{(2)},\ \cdots\ \boldsymbol{e}_j^{(k)}\right\}$ is linearly independent.

PROOF: (We did not give a proof for the case of distinct eigenvalues, the approach used here can be employed in other cases.) Suppose that $\left\{\boldsymbol{e}_j^{(1)},\ \boldsymbol{e}_j^{(2)},\ \cdots\ \boldsymbol{e}_j^{(k)}\right\}$ are linearly dependent, then there exist constants $c_1,\ c_2,\ \cdots\ c_k \in C$ which are not all zero such that

$$\sum_{i=1}^{k} \alpha_i\, \boldsymbol{e}_j^{(i)} = \mathbf{0}. \tag{8.67}$$

Multiplying $(A-\lambda_j I)^{k-1}$ to (8.67) and observing that

$$(A-\lambda_j I)^{k-1}\boldsymbol{e}_j^{(i)} = (A-\lambda_j I)^{2k-(i+1)}\boldsymbol{e}_j = \mathbf{0} \quad \text{for} \quad i=1\,,\ 2\,,\cdots,\ k-1 \tag{8.68}$$

one obtains

$$\alpha_k\,(A-\lambda_j I)^{k-1}\boldsymbol{e}_j^{(k)} = \mathbf{0} \tag{8.69}$$

which, in turn, results in $\alpha_k = 0$ due to $(A-\lambda_j I)^{k-1}\boldsymbol{e}_j \neq \mathbf{0}$. Next, multiplying $(A-\lambda_j I)^{k-2}$ yields $\alpha_{k-1} = 0$, etc.

(ii) The generalized eigenvectors of A associated with different eigenvalues are linearly independent.

(iii) The above approach to find the generalized eigenvectors of A may be called a *backward* procedure since one finds $\boldsymbol{e}_j^{(k-1)}$ from $\boldsymbol{e}_j^{(k)}$. An alternative, which may be called a *forward* procedure, can also be used, i.e. deriving $\boldsymbol{e}_j^{(k)}$ from $\boldsymbol{e}_j^{(k-1)}$. It depends upon how you define the generalized eigenvectors. The forward procedure is described as follows. First note that

$$(A-\lambda_j I)\,\boldsymbol{e}_j^{(1)} = (A-\lambda_j I)\,(A-\lambda_j I)^{k-1}\boldsymbol{e}_j = (A-\lambda_j I)^k\,\boldsymbol{e}_j = \mathbf{0} \tag{8.70}$$

where the last step holds due to the definition of generalized eigenvectors. Therefore, having found $\boldsymbol{e}_j^{(1)}$ from

$$(A-\lambda_j I)\,\boldsymbol{e}_j^{(1)} = \mathbf{0}$$

one can find other generalized eigenvectors using the following linear algebraic equations:

$$\begin{aligned} A-\lambda_j I)\,\boldsymbol{e}_j^{(2)} &= \boldsymbol{e}_j^{(1)} \\ &\vdots \\ A-\lambda_j I)\,\boldsymbol{e}_j^{(k)} &= \boldsymbol{e}_j^{(k-1)}. \end{aligned} \tag{8.71}$$

Which procedure is better? The backward procedure is better than the forward procedure since the former needs multiplication operations only while the latter also requires solving linear algebraic equations.

EXAMPLE 8.5 (continued). Let $\boldsymbol{e}_1$ be a vector such that

$$(A - \lambda_1 I)\,\boldsymbol{e}_1 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix} \boldsymbol{e}_1 \neq \mathbf{0} \tag{8.72}$$

but

$$(A - \lambda_1 I)^2\,\boldsymbol{e}_1 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix} \boldsymbol{e}_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & -3 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{e}_1 = \mathbf{0}. \tag{8.73}$$

Clearly, $\boldsymbol{e}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ satisfies these conditions. It is a generalized eigenvector of rank 2. Let

$$\boldsymbol{e}_1^{(2)} = \boldsymbol{e}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{8.74}$$

then

$$\boldsymbol{e}_1^{(1)} = (A - \lambda_1 I)\,\boldsymbol{e}_1^{(2)} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{8.75}$$

We have found $\boldsymbol{e}_3 = \begin{pmatrix} 1 \\ -3 \\ 1 \end{pmatrix}$. From the above remarks, we know that $\boldsymbol{e}_1^{(1)}$, $\boldsymbol{e}_1^{(2)}$ and $\boldsymbol{e}_3$ are linearly independent, so we can now construct the transformation matrix $T = (\boldsymbol{e}_1^{(1)} \;\; \boldsymbol{e}_1^{(2)} \;\; \boldsymbol{e}_3) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}$, and thus, $T^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$.

Finally, we obtain

$$\begin{aligned} \bar{A} &= T^{-1} A T \\ &= \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 2 \\ 0 & -1 & 3 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 & 4 \\ 0 & -1 & -3 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \end{aligned} \tag{8.76}$$

which is in the Jordan canonical form. A more complicated example is given next.

EXAMPLE 8.6. $A = \begin{bmatrix} 3 & -1 & -1 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$.

Answer:

$$T = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & -2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\bar{A} = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

8.3 Stability of $\boldsymbol{x} = 0$ in the Associated Linear System About Fixed Points

Now, let's return to the discussion of the stability for $\boldsymbol{x} = \mathbf{0}$ of the differential equation $\dot{\boldsymbol{x}} = A\,\boldsymbol{x}$. Here it is assumed that A is a real matrix because physical systems usually generate real matrices. First, consider A given in Example 8.3, which has eigenvalues $\lambda_1 = -1 + i$, $\lambda_2 = -1 - i$ and $\lambda_3 = -1$. They are distinct and all the real parts of the eigenvalues are negative, so $\boldsymbol{x} = \mathbf{0}$ is stable. In fact, the solution

$$\begin{aligned} \boldsymbol{x} &= e^{At}\,\boldsymbol{x}_0 = e^{T\bar{A}T^{-1}t}\,\boldsymbol{x}_0 = Te^{\bar{A}t}\,T^{-1}\,\boldsymbol{x}_0 \\ &= \begin{bmatrix} 1-2\,i & 1+2\,i & 1 \\ 1 & 1 & 0 \\ i & -i & 0 \end{bmatrix} \begin{bmatrix} e^{(-1+i)t} & 0 & 0 \\ 0 & e^{(-1-i)t} & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \frac{1}{2} \begin{bmatrix} 0 & -1 & -i \\ 0 & 1 & i \\ 2 & -2 & 4 \end{bmatrix} \boldsymbol{x}_0 \end{aligned}$$

$$= e^{-t}\begin{bmatrix} (1-2i)e^{it} & (1+2i)e^{-it} & 1 \\ e^{it} & e^{-it} & 0 \\ i\,e^{it} & -i\,e^{-it} & 0 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 0 & -1 & -i \\ 0 & 1 & i \\ 2 & -2 & 4 \end{bmatrix}\boldsymbol{x}_0$$

$$= \frac{1}{2}e^{-t}\begin{bmatrix} 2 & (1-2\,i)\,e^{it}+(1+2\,i)\,e^{-it}-2 & (-2-i)\,e^{it}+(-2+i)\,e^{-it}+4 \\ 0 & e^{it}+e^{-it} & -i\,e^{it}+i\,e^{-it} \\ 0 & i\,e^{it}-i\,e^{-it} & e^{it}+e^{-it} \end{bmatrix}\boldsymbol{x}_0$$

$$= e^{-t}\begin{bmatrix} 1 & \cos t+2\sin t-1 & \sin t-2\cos t+2 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{bmatrix}\boldsymbol{x}_0$$

where the **Euler**'s formulas $\cos t = \frac{1}{2}(e^{it}+e^{-it})$ and $\sin t = \frac{1}{2i}(e^{it}-e^{-it})$ have been used. It is clearly to see that $\|\boldsymbol{x}\|$ (or all the components of $\boldsymbol{x}$) $\to 0$ when $t \to \infty$ due to $e^{-t} \to 0$. The above results can also be obtained from the real transform as follows:

$$\boldsymbol{x} = e^{At}\boldsymbol{x}_0 = e^{T\bar{A}T^{-1}t}\boldsymbol{x}_0 = Te^{\bar{A}t}T^{-1}\boldsymbol{x}_0$$

$$= \begin{bmatrix} 1 & -2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} e^{-t}\cos t & e^{-t}\sin t & 0 \\ -e^{-t}\sin t & e^{-t}\cos t & 0 \\ 0 & 0 & e^{-t} \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 2 \end{bmatrix}\boldsymbol{x}_0$$

$$= e^{-t}\begin{bmatrix} \cos t+2\sin t & \sin t-2\cos t & 1 \\ \cos t & \sin t & 0 \\ -\sin t & \cos t & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 2 \end{bmatrix}\boldsymbol{x}_0$$

$$= e^{-t}\begin{bmatrix} 1 & \cos t+2\sin t-1 & \sin t-2\cos t+2 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{bmatrix}\boldsymbol{x}_0$$

where the formula $\exp\left(\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}t\right) = \begin{bmatrix} e^{-t}\cos t & e^{-t}\sin t \\ -e^{-t}\sin t & e^{-t}\cos t \end{bmatrix}$ has been used, which can be proved by direct Taylor expansions.

The similar computation can be applied to Example 8.4 and find that the system is unstable because one eigenvalue is positive. Finally, consider Example 8.5, it is seen that all the real parts of the eigenvalues are negative, but since one of the eigenvalue is repeated, we cannot conclude stability directly. Let's compute $\boldsymbol{x}$

$$\boldsymbol{x} = e^{At}\boldsymbol{x}_0 = e^{T\bar{A}T^{-1}t}\boldsymbol{x}_0 = Te^{\bar{A}t}T^{-1}\boldsymbol{x}_0$$

$$
\begin{aligned}
&= T \exp \begin{bmatrix} -t & t & 0 \\ 0 & -t & 0 \\ 0 & 0 & -2t \end{bmatrix} T^{-1} \boldsymbol{x}_0 \\
&= T \begin{bmatrix} 1 - t + \frac{1}{2!}t^2 & t - \frac{1}{1!}t^2 + \frac{1}{2!}t^3 & 0 \\ 0 & 1 - t + \frac{1}{2!}t^2 & 0 \\ 0 & 0 & 1 - 2\,t + \frac{1}{2!}(2t)^2 \end{bmatrix} T^{-1} \boldsymbol{x}_0 \\
&= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e^{-t} & te^{-t} & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x}_0 \\
&= \begin{bmatrix} e^{-t} & t\,e^{-t} & e^{-2t} \\ 0 & e^{-t} & -3\,e^{-2t} \\ 0 & 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix} \boldsymbol{x}_0 \\
&= \begin{bmatrix} e^{-t} & t\,e^{-t} & e^{-t} + 3\,t\,e^{-t} + e^{-2t} \\ 0 & e^{-t} & 3\,e^{-t} - 3\,e^{-2t} \\ 0 & 0 & e^{-2t} \end{bmatrix} \boldsymbol{x}_0
\end{aligned}
$$

which shows that $\|\boldsymbol{x}\|$ (or all the components of $\boldsymbol{x}$) $\to 0$ when $t \to \infty$ because $e^{-t} \to 0$, $e^{-2t} \to 0$ and $t\,e^{-t} \to 0$ when $t \to \infty$. In general, for a repeated eigenvalue λ_j of rank k of A, the solution $\boldsymbol{x}$ corresponding to this part is in the form of

$$
e^{\lambda_j t} \left(\boldsymbol{e}_j^{(k)} + \frac{1}{2!}\boldsymbol{e}_j^{(k-1)} t + \frac{1}{3!}\boldsymbol{e}_j^{(k-2)} t^2 + \cdots + \frac{1}{k!}\boldsymbol{e}_j^{(1)} t^{k-1} \right),
$$

where $\boldsymbol{e}_j^{(i)}$ $i = 1, 2, \cdots, k$ are the associated generalized eigenvectors. Thus, the conclusion is that the λ's will still decide stability since polynomials cannot compare to exponentials in this regard. Therefore, we have the following conclusion for the **stability of $\boldsymbol{x} = \mathbf{0}$ of the associated linear system $\dot{\boldsymbol{x}} = A\,\boldsymbol{x}$** :

1. If all the real parts of the eigenvalues of A are negative, then $\boldsymbol{x} = \mathbf{0}$ is (asymptotically) stable.

2. If at least one of the eigenvalues has positive real part, then $\boldsymbol{x} = \mathbf{0}$ is unstable.

3. If some of the eigenvalues have zero real parts, and the remaining have negative real

parts, then the linear approximation fails to determine the stability of $\boldsymbol{x} = \mathbf{0}$, which is called **critical** cases. It needs nonlinear analysis, including higher order nonlinear terms, to determine the stability.

8.4 Stability and Phase Portraits Near Fixed Points in the Phase Plane

In the previous section, we discussed the stability of $\boldsymbol{x} = \mathbf{0}$ of the associated linearized system of a nonlinear DE. Especially, we considered fixed points. Now, we want to consider stability of solutions of the DE near fixed points – phase portraits in phase plane.

It is fairly clear that stability is decided by the real part of the eigenvalues in general, for the linear case. We'll extend this to the nonlinear case shortly. In the phase plane (two dim. space) we can examine how trajectories behave near the origin based upon this. This proves to be helpful in subsequent discussions.

Consider

$$\dot{\boldsymbol{x}} = A\,\boldsymbol{x} \qquad \text{or} \qquad \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{8.77}$$

The eigenvalues of A come from

$$\det(\lambda\, I - A) = \det = \begin{bmatrix} \lambda - a_{11} & -\,a_{12} \\ -\,a_{21} & \lambda - a_{22} \end{bmatrix} = 0$$

which results in

$$\lambda^2 - p\,\lambda + q = 0, \tag{8.78}$$

and so we have eigenvalues

$$\begin{aligned} \lambda_1 &= \frac{1}{2}\,(p - \sqrt{p^2 - 4\,q}) = \frac{1}{2}\,(p - \sqrt{\Delta}) \\ \lambda_2 &= \frac{1}{2}\,(p + \sqrt{p^2 - 4\,q}) = \frac{1}{2}\,(p + \sqrt{\Delta}) \end{aligned} \tag{8.79}$$

where

$$\begin{aligned} p &= a_{11} + a_{22} && \text{trace} \\ q &= a_{11}\,a_{22} - a_{12}\,a_{21} && \text{determinant} \\ \Delta &= p^2 - 4\,q && \text{discriminant}\,. \end{aligned} \tag{8.80}$$

Based on the eigenvalues λ_1 and λ_2 (or the coefficients a_{ij}), there exist seven cases, considered below.

Case I: Saddle. when λ_1 and λ_2 are both real and have opposite signs, i.e. $\lambda_1 < 0 < \lambda_2$ $(q < 0)$.

A suitable change of the coordinates $\boldsymbol{x} = P\,\boldsymbol{y}$ puts the equation $\dot{\boldsymbol{x}} = A\,\boldsymbol{x}$ in the form $\dot{\boldsymbol{y}} = B\,\boldsymbol{y}$ where $B = P^{-1}\,A\,P = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. Thus, the phase portrait of the new system looks like Fig. 7.3, and the phase portrait of the original system is depicted in Fig. 8.1.

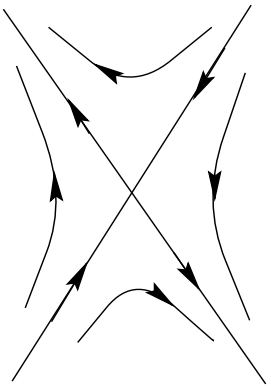


Fig. 8.1 Case I: saddle.

Case II: Node. when λ_1 and λ_2 are distinct, non-zero and have same signs, i.e. either (i) **stable node** if $\lambda_1 < \lambda_2 < 0$ $(p < 0)$ or (ii) **unstable node** if $0 < \lambda_1 < \lambda_2$ $(p > 0)$. This case is demonstrated in Fig. 8.2.

Case III: Focus (spiral). when λ_1 and λ_2 are a pair of complex conjugates, i.e. $\lambda_{1,2} = a \pm i\,b$ where $a = \dfrac{1}{2}\,p$ and $b = \dfrac{1}{2}\,\sqrt{-\Delta} > 0$ $(p \neq 0\,,\ \Delta < 0)$. Then one can change the coordinates to obtain the equivalent system

$$\dot{\boldsymbol{y}} = B\,\boldsymbol{y}\,, \qquad B = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \tag{8.81}$$

and

$$e^{t\,B} = e^{t\,a} \begin{bmatrix} \cos tb & \sin tb \\ -\sin tb & \cos tb \end{bmatrix}. \tag{8.82}$$

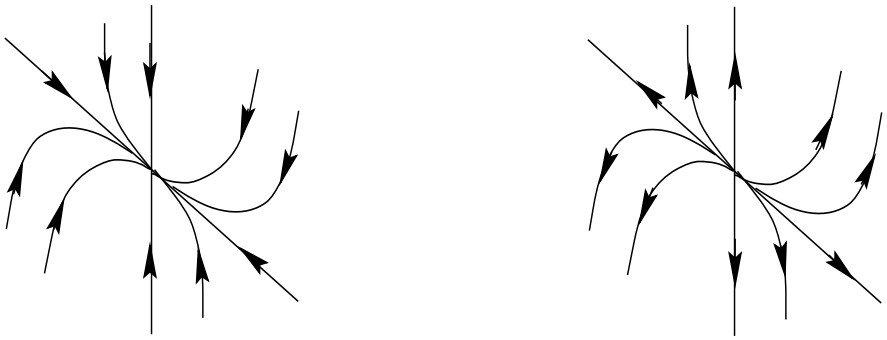


Fig. 8.2 Case II: node: (i) stable node; (ii) unstable node.

Thus, the solution is given by

$$\boldsymbol{y} = e^{t\,B}\boldsymbol{k} = e^{t\,a}\,(k_1\,\cos\,tb + k_2\,\sin\,tb\,,\;\; k_2\,\cos\,tb - k_1\,\sin\,tb)\,. \tag{8.83}$$

There are two subcases: (i) **stable focus** if $\mathrm{Re}(\lambda\text{'s}) < 0$, i.e. $p < 0$; (ii) **unstable focus** if $\mathrm{Re}(\lambda\text{'s}) > 0$, i.e. $p > 0$. This case is shown in Fig. 8.3. If the matrix B in equation (8.81) is chosen as $B = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, then how should you change Fig. 8.3?

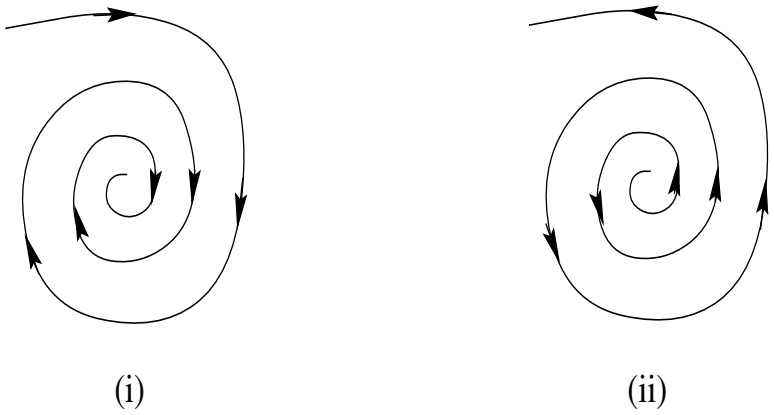


Fig. 8.3 Case III: focus: (i) stable focus; (ii) unstable focus.

The three cases discussed above are **robust** cases since a small perturbation of the system (in terms of changing a_{ij}) does not change the type of trajectories (or stability which will be discussed in the next chapter). The next four cases are **non-robust** cases.

Case IV: when one of λ_1 and λ_2 is equal to zero. Without loss of generality, assume that $\lambda_1 = 0$. Then we have either (i) $\lambda_2 = p < 0$ or (ii) $\lambda_2 = p > 0$. This case is shown in Fig. 8.4.

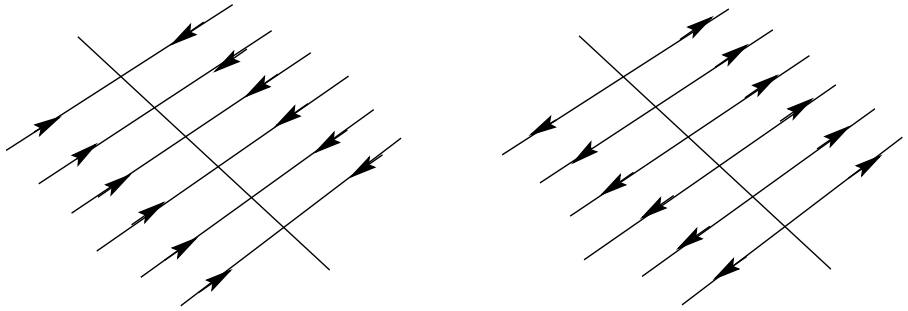


Fig. 8.4 Case IV: $\lambda_1 = 0$: (i) $\lambda_2 < 0$; (ii) $\lambda_2 > 0$.

Case V: when $\lambda_1 = \lambda_2 = \frac{1}{2}p$ $(\Delta = 0)$.

(i) $a_{12} = a_{21} = 0$ $(\Delta = 0 \Longleftrightarrow a_{11} = a_{22})$. In this case, A is of *simple* structure and it can be diagonalized. There are two subcases: (a) **degenerate stable node** if $p < 0$; (b) **degenerate unstable node** if $p > 0$. A degenerate node is also called an *improper* node. This subcase is depicted in Fig. 8.5.

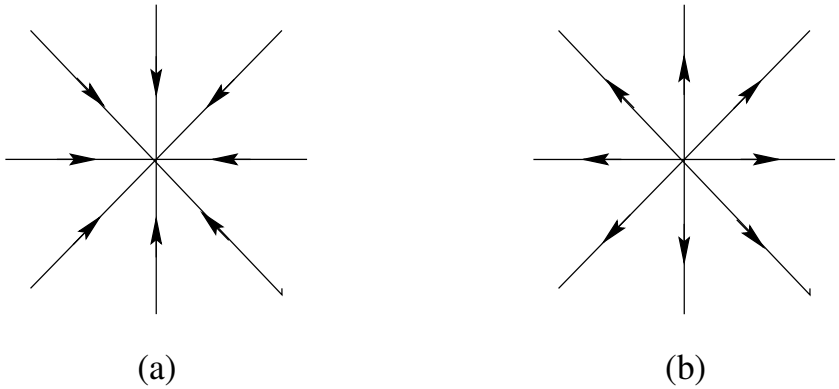


Fig. 8.5 Case V(i): (a) degenerate stable node; (b) degenerate unstable node.

(ii) $a_{12}\, a_{21} \neq 0$. In this case, A does not have a simple structure and cannot be diago-

nalized. There is a change of coordinates $\boldsymbol{x} = P\,\boldsymbol{y}$ giving the equivalent equation

$$\dot{\boldsymbol{y}} = B\,\boldsymbol{y}\ , \qquad B = P^{-1}\,A\,P = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \tag{8.84}$$

and the solutions look like Fig. 8.6.

Fig. 8.6 Case V(ii): $a_{12}\,a_{21} \neq 0$: (a) $p < 0$; (b) $p > 0$.

Case VI: Center. when $\lambda_{1,2} = \pm\, i\,\omega$ where $\omega = \sqrt{q}$ $(p = 0\,,\ q > 0)$. Now the solutions of the system are periodic with the same period. To see this, change coordinates to get

$$\dot{\boldsymbol{y}} = B\,\boldsymbol{y}\ , \qquad B = \begin{bmatrix} 0 & b \\ -\,b & 0 \end{bmatrix}. \tag{8.85}$$

We know that

$$e^{t\,B} = \begin{bmatrix} \cos tb & \sin tb \\ -\sin tb & \cos tb \end{bmatrix}. \tag{8.86}$$

Hence, $\boldsymbol{y}(t + \dfrac{2\pi}{b}) = \boldsymbol{y}(t)$. So in the $\boldsymbol{y}$ coordinates the phase portrait consists of concentric circles, which will be ellipses in $\boldsymbol{x}$ coordinates (see Fig. 8.7).

Case VII: when $\lambda_1 = \lambda_2 = 0$ $(p = q = 0)$. This is shown in Fig. 8.8.

The four **non-robust** cases (IV)–(VII) represent borders between the robust cases; that is, any change no matter how small in the components of A can cause them to change to another case. (This is referred as *structural instability.*)

All the seven cases discussed above can be summarized in a single graph (see Fig. 8.9). Note that case (V) is not sensitive regarding the case of stability, only in the case of real to complex. Thus, the only border line cases in terms of stability, are along the p-axis, and positive q-axis. So, *(asymptotic) stability is in the second quadrant only.*

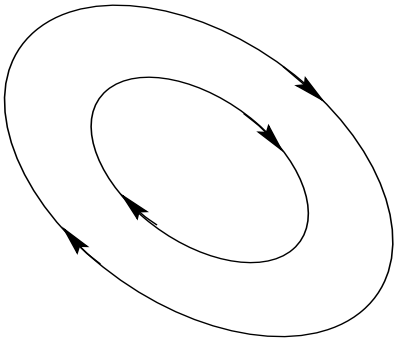


Fig. 8.7 Case VI: center.

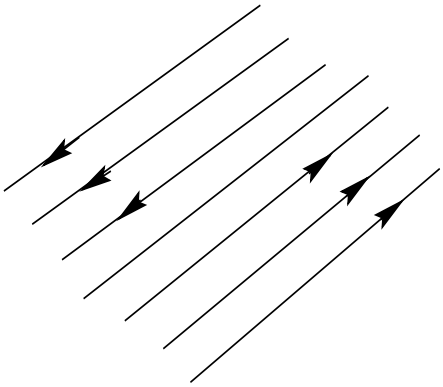


Fig. 8.8 Case VII: $\lambda_1 = \lambda_2 = 0$.

It should be pointed out that the above classification of seven cases may be given in different versions. For example (cf. Hirsch & Smale), one can use a *four* case classification: (1) A has real eigenvalues of opposite signs (**Saddle**); (2) All eigenvalues of A have negative real parts (**Sink**); (3) All eigenvalues of A have positive real parts (**Source**); and (4) The eigenvalues are pure imaginary (**Center**).

Now let's consider an example to see how to obtain those trajectories shown in Figs 8.1 – 8.8, given in the original coordinates $\boldsymbol{x}$. Choose the saddle case, for simplicity, which has non-zero real eigenvalues of opposite signs.

EXAMPLE 8.7. Consider the system $\dot{\boldsymbol{x}} = A\,\boldsymbol{x}$ where A has non-zero real eigenvalues of opposite signs at the equilibrium $\boldsymbol{x} = \boldsymbol{0}$. Draw the phase portrait of the system near the origin $\boldsymbol{x} = \boldsymbol{0}$.

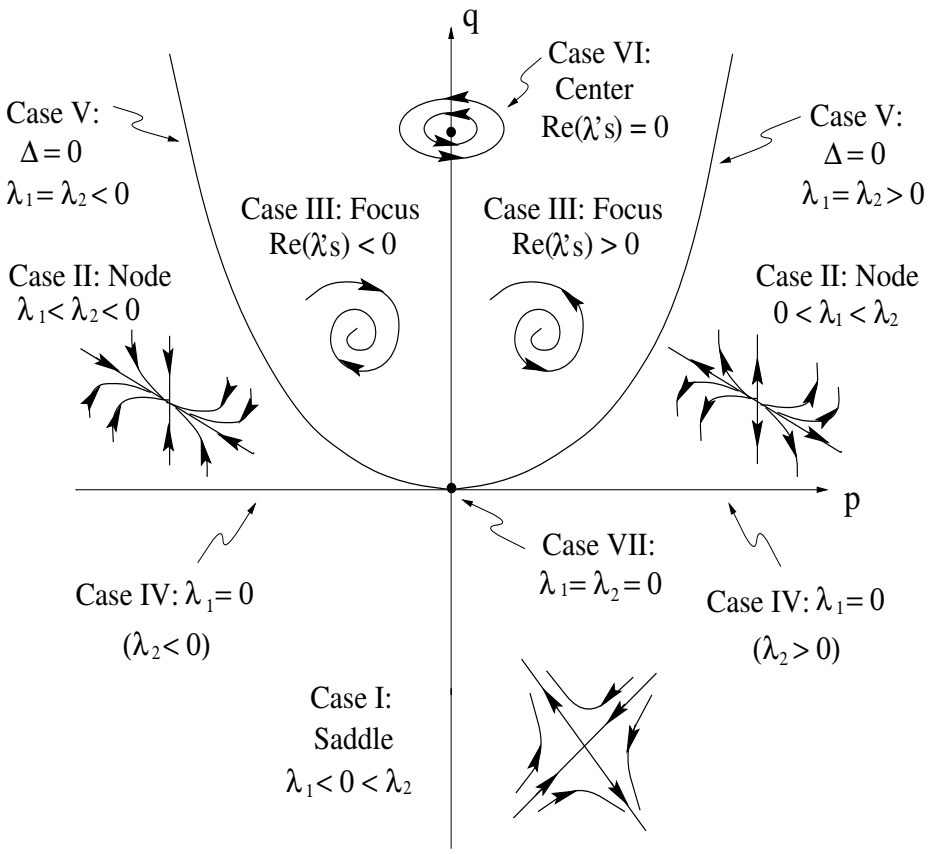


Fig. 8.9 Seven different cases for classifying qualitative behavior of a 2-dim. linear D.E. system

Solution: By introducing the transformation $\boldsymbol{x} = P\,\boldsymbol{y}$, and using the solution $\boldsymbol{x} = e^{t\,A}\,\boldsymbol{k}$ and the relation $A = P\,B\,P^{-1}$ where $B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, we can obtain

$$\begin{aligned} \boldsymbol{x} &= P \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} P^{-1}\,\boldsymbol{k} \\ &= (\boldsymbol{v}_1 \;\; \boldsymbol{v}_2) \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} P^{-1}\,\boldsymbol{k} \\ &= (\boldsymbol{v}_1\, e^{\lambda_1 t} \;\; \boldsymbol{v}_2\, e^{\lambda_2 t})\,\boldsymbol{C} \\ &= C_1\, e^{\lambda_1 t}\,\boldsymbol{v}_1 + C_2\, e^{\lambda_2 t}\,\boldsymbol{v}_2 \end{aligned} \tag{8.87}$$

where $\boldsymbol{C} = P^{-1}\,\boldsymbol{k} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$ is an arbitrary vector. Thus, solutions with $C_2 = 0$, are $C_1\,\boldsymbol{v}_1\,e^{\lambda_1 t} = C_1 \begin{pmatrix} \alpha \\ \beta \end{pmatrix} e^{\lambda_1 t}$, or

$$x_1 = C_1\,\alpha\,e^{\lambda_1 t} \qquad \text{and} \qquad x_2 = C_1\,\beta\,e^{\lambda_1 t}\,. \tag{8.88}$$

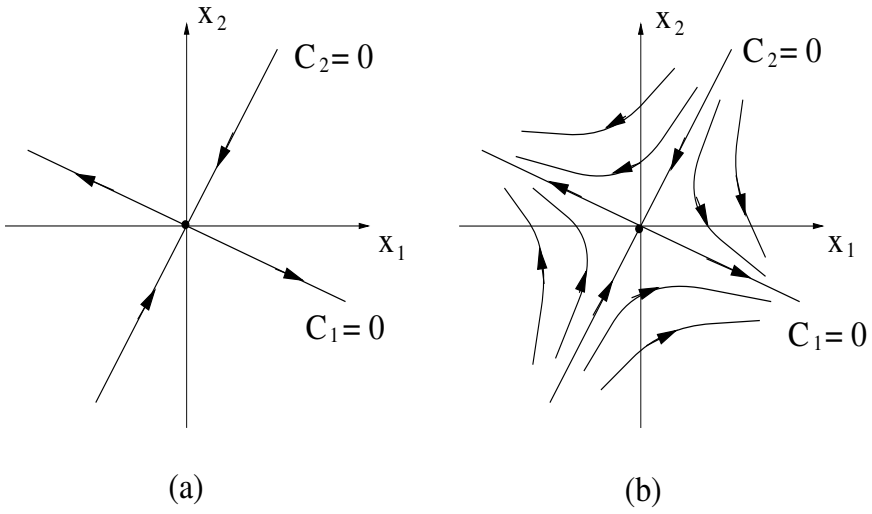


Fig. 8.10 Trajectories of saddle case: (a) two particular trajectories; (b) phase portrait.

Thus, $\dfrac{x_2}{x_1} = \dfrac{\beta}{\alpha}$, or $x_2 = (\text{const.})\, x_1$ which represents a straight line through the origin. Similarly, for $C_1 = 0$, we have another straight line. Note that the slopes of these two lines are actually the *eigenvectors* corresponding to the eigenvalues λ_1 and λ_2, respectively. Since $\lambda_1 < 0$ and $\lambda_2 > 0$ for the saddle case, x_1 and x_2 shrink along the line $C_2 = 0$ while they grow along the line $C_1 = 0$ (see Fig. 8.10(a)). Then the adjacent trajectories can be easily obtained as in Fig. 8.10(b).

Let's consider another example.

EXAMPLE 8.8. Again consider the linear D.E. $\dot{\boldsymbol{x}} = A\,\boldsymbol{x}$ where $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. Define $\Delta \equiv p^2 - 4\,q$ where $p = a_{11} + a_{22}$ and $q = a_{11}\,a_{22} - a_{12}\,a_{21}$. Make phase portraits of the system near the origin for the case: $\Delta > 0$.

Solution: When $\Delta > 0$, both the two eigenvalues are real and distinct. The phase portraits are either **saddle** or **node**. The saddle case has been detailed discussed in Example 8.7. So here we only discuss the case of node only. In fact, the discussion of node case can follow that of saddle case. By using the equation (8.87) we have the solution $\boldsymbol{x} = C_1\,\boldsymbol{v}_1\,e^{\lambda_1 t} + C_2\,\boldsymbol{v}_2\,e^{\lambda_2 t}$ where C_1 and C_2 are arbitrary constants, and $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are eigenvectors corresponding to eigenvalues λ_1 and λ_2, respectively. Suppose that $\lambda_2 > \lambda_1 > 0$ which implies $p > 0$. (Note: reversing the signs of the λ's means contraction, i.e. reversing the arrows. $p < 0 \Longrightarrow$

$\lambda_1 < \lambda_2 < 0$.) In this case, as $t \to -\infty$, both $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ shrink, so x_1 and x_2 shrink to zero as $t \to -\infty$. Let $\boldsymbol{v}_i = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}$ for $i = 1, 2$. Then

$$\begin{aligned} \frac{dx_2}{dx_1} &= \frac{\dfrac{dx_2}{dt}}{\dfrac{dx_1}{dt}} = \frac{C_1 \beta_1 \lambda_1 e^{\lambda_1 t} + C_2 \beta_2 \lambda_2 e^{\lambda_2 t}}{C_1 \alpha_1 \lambda_1 e^{\lambda_1 t} + C_2 \alpha_2 \lambda_2 e^{\lambda_2 t}} \\ &\quad \text{(which is } \frac{0}{0} \text{ indeterminant form as t} \to -\infty) \\ &= \frac{C_1 \beta_1 \lambda_1 + C_2 \beta_2 \lambda_2 e^{(\lambda_2 - \lambda_1) t}}{C_1 \alpha_1 \lambda_1 + C_2 \alpha_2 \lambda_2 e^{(\lambda_2 - \lambda_1) t}} \\ &= \frac{\beta_1}{\alpha_1} \qquad \text{as} \quad t \to -\infty . \end{aligned} \tag{8.89}$$

which is the slope of the straight line represented by $C_2 = 0$, corresponding to the "slow" eigendirection ($\boldsymbol{v}_1$). Another line is given by $C_1 = 0$ which has the slope $dx_2/dx_1 = \beta_2/\alpha_2$, corresponding to the "fast" eigendirection ($\boldsymbol{v}_2$). Having found these two lines, the adjacent trajectories can be easily determined. The phase portraits are shown in Fig. 8.11.

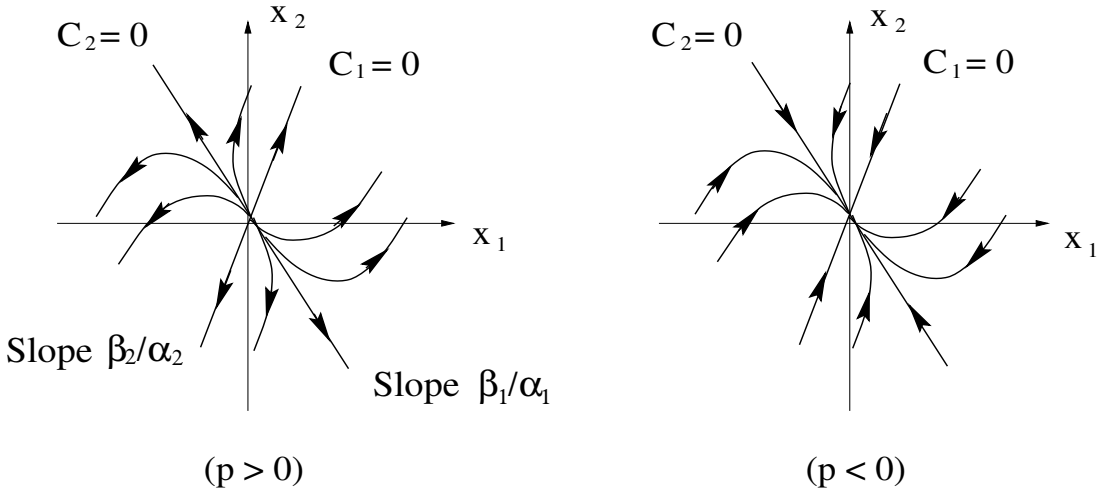


Fig. 8.11 Phase portrait for Example 8.8.

Particular cases: if $p = 0$, then $\Delta = -4q$ $(q < 0)$ and thus $\lambda_{1,2} = \pm\sqrt{-q}$ which is a saddle case. If $q = 0$, then one of the eigenvalues is 0. Let $\lambda_1 = 0$, so $\lambda_2 = p$. Hence, the solutions are given by

$$\begin{aligned} x_1 &= C_1 \alpha_1 + C_2 \alpha_2 e^{\lambda_2 t} \\ x_2 &= C_1 \beta_1 + C_2 \beta_2 e^{\lambda_2 t} \end{aligned} \tag{8.90}$$

which, in turn, result in

$$\begin{aligned} x_1(t) &\longrightarrow C_1\,\alpha_1 \quad \text{and} \quad x_2(t) \longrightarrow C_1\,\beta_1 \quad \text{as} \quad t \longrightarrow \infty \quad \text{if} \quad p<0 \\ x_1(t) &\longrightarrow C_1\,\alpha_1 \quad \text{and} \quad x_2(t) \longrightarrow C_1\,\beta_1 \quad \text{as} \quad t \longrightarrow -\infty \quad \text{if} \quad p>0. \end{aligned} \tag{8.91}$$

It can be seen from (8.91) that all the trajectories either shrink ($p<0$) to or grow ($p>0$) from the line

$$x_2 = \frac{\beta_1}{\alpha_1}\,x_1\,. \tag{8.92}$$

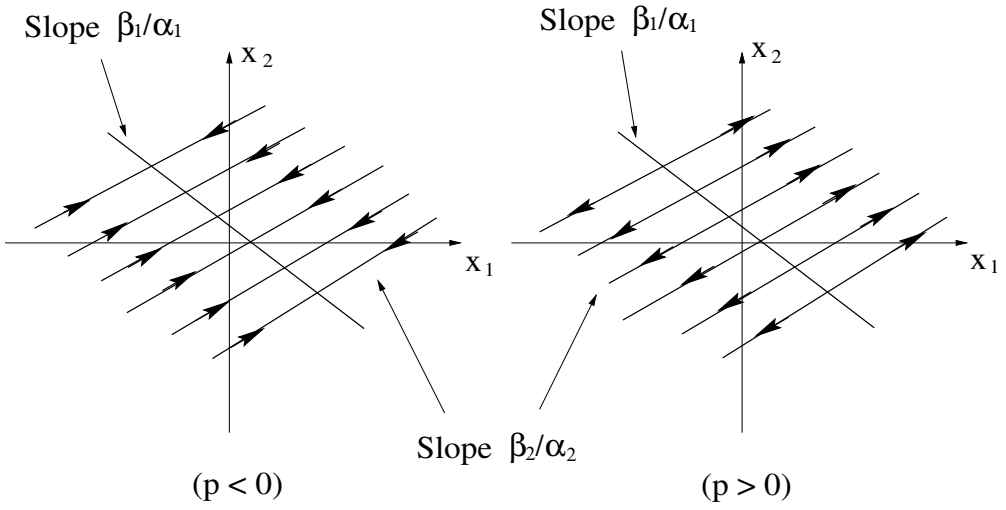


Fig. 8.12 Phase portrait for Example 8.8 when $q=0$ ($\Delta>0$).

Moreover, (8.90) yields a family of parallel lines, given by

$$x_2 = \frac{\beta_2}{\alpha_2}\,(x_1 - C_1\,\alpha_1) - C_1\,\beta_1 = (\frac{\beta_2}{\alpha_2})\,x_1 - C_1\,\left[\,\beta_1 + (\frac{\beta_2}{\beta_1})\,\alpha_1\,\right] \tag{8.93}$$

in which only C_1 is an arbitrary constant and the slope of the parallel lines is $\dfrac{\beta_2}{\alpha_2}$. The phase portrait is given in Fig. 8.12.

8.5 What Is Stability and What Does Linear Stability Have to Do with It?

In section 8.1 we asked two questions, essentially,

1. What makes linear stability?

2. What does it have to do with stability in general (to the original DE)?

The first question we have answered in the previous sections. We now turn to the second more important question. The problem we face immediately is what is stability anyway. We need some definitions, and must recognize that the notion of stability ultimately induces a variety of definitions.

Loosely speaking, stability of $\tilde{\boldsymbol{y}}(t)$, which is a solution of the DE $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y})$, concerns closeness, as we understand. If any $\boldsymbol{y}(t)$ starts close to $\tilde{\boldsymbol{y}}(t)$ remains close to $\tilde{\boldsymbol{y}}(t)$, then $\tilde{\boldsymbol{y}}(t)$ is stable. It is **asymptotically** stable if nearby solutions actually converge to $\tilde{\boldsymbol{y}}(t)$ as $t \to \infty$. In fact, the stability we discussed in the previous sections is asymptotic stability.

Lagrange stability only refers to $\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t) = \boldsymbol{x}(t)$ being bounded: a generous form of stability, which may nonetheless be the heart of the matter. In this sense "close" means simply non-infinite.

Other stability definitions such as **Liapunov stability**, **Input-Output stability**, etc. are frequently used. Stability is a big topic, and we will only discuss Liapunov stability in detail. Liapunov stability is something in between asymptotic and Lagrange stability: $\tilde{\boldsymbol{y}}(t)$ is said to be "L" stable if, given $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$, for any other solutions $\boldsymbol{y}(t)$ such that

$$\|\boldsymbol{y}(t_0) - \tilde{\boldsymbol{y}}(t_0)\| < \delta \tag{8.94}$$

then,

$$\|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\| < \epsilon \qquad \forall\ t > t_0\,. \tag{8.95}$$

Remark: This allows for the possible growth of $\|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\|$. The point is that in principle any final deviation, ϵ, no matter how small, can be selected by choosing a small enough initial deviation, δ, which may be greater or less than ϵ.

8.5.1. Stability of a Fixed Point (an Equilibrium State). Suppose that the differential equation is described by the autonomous system

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}) \tag{8.96}$$

and that there exists $\boldsymbol{y}_e$ such that $\boldsymbol{f}(\boldsymbol{y}_e) = \boldsymbol{0}$; i.e., $\boldsymbol{y}_e$ is a fixed point (equilibrium). Apply

the transformation $\boldsymbol{y} = \boldsymbol{y}_e + \boldsymbol{x}$ into the above DE to obtain the associated linear system

$$\dot{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{y}_e)\,\boldsymbol{x} = A\,\boldsymbol{x} \tag{8.97}$$

where $\boldsymbol{x} = \boldsymbol{0}$ is a fixed point, corresponding to the solution $\boldsymbol{y}_e$ of the original DE.

"L" stability: $\boldsymbol{x} = \boldsymbol{0}$ is stable if it is possible to find a $\delta(\epsilon) > 0$ for any given $\epsilon > 0$ such that the motion following the initial disturbance

$$\|\boldsymbol{x}_0\| < \delta \qquad \text{where} \qquad \boldsymbol{x}_0 = \boldsymbol{x}(0) \tag{8.98}$$

satisfies

$$\|\boldsymbol{x}(t)\| < \epsilon \qquad \forall \quad t > 0\,. \tag{8.99}$$

Otherwise, the fixed point $\boldsymbol{x} = \boldsymbol{0}$ is unstable. $\boldsymbol{x} = \boldsymbol{0}$ is said to be asymptotically stable if it is stable and, in addition,

$$\lim_{t\to\infty} \|\boldsymbol{x}(t)\| = 0. \tag{8.100}$$

These definitions imply that the fixed point (equilibrium state) represented by the origin of $\boldsymbol{R}^n$ is stable if the variable point $\boldsymbol{y}(t)$, describing the trajectories of the system in $\boldsymbol{R}^n$, remains within a hypersphere of arbitrarily small radius ϵ, provided the initial disturbance is such that the point $\boldsymbol{y}_e$ is within a hypersphere of sufficiently small radius δ. If the point $\boldsymbol{y}(t)$ also tends to the origin as $t \to \infty$, then the origin is asymptotically stable. On the other hand, the origin is unstable if the trajectory of the point $\boldsymbol{y}(t)$ crosses the hypersphere of radius ϵ. A two-dimensional representation of these phenomena is depicted in Fig. 8.13.

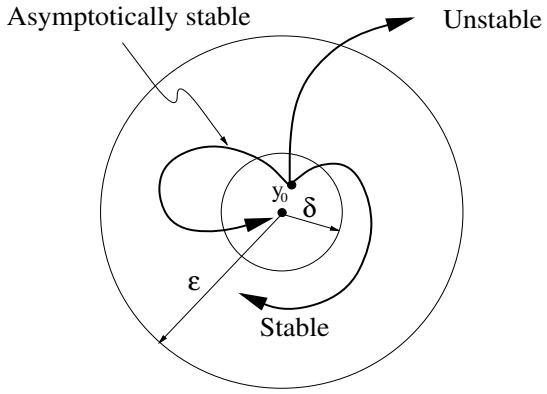


Fig. 8.13 Stability concept for an autonomous system.

$\boldsymbol{x} = \mathbf{0}$ is said to be globally asymptotically stable if it is stable and, in addition, every motion converges to $\boldsymbol{x} = \mathbf{0}$ as $t \to \infty$. This property is also referred to as asymptotic stability in the large.

EXAMPLE 8.9. Suppose the table top is shown as in Fig. 8.14, and assume that there is a frictional force proportional to $\dot{r}$. Then (exercise):

(i) $r = r_1\,,\ \dot{r} = 0:$ asymptotically stable;

(ii) $r = r_2\,,\ \dot{r} = 0:$ unstable;

(iii) $r = r_3\,,\ \dot{r} = 0:$ stable.

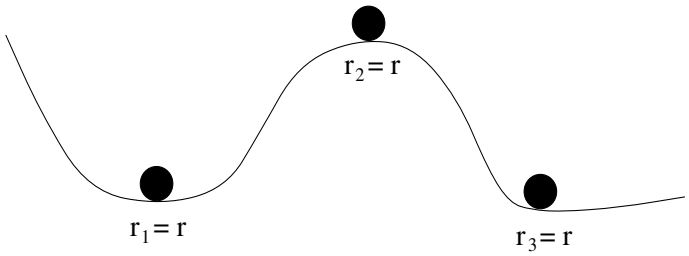


Fig. 8.14 Figure of Example 8.9.

Applying the stability definitions to study stability of an equilibrium is not convenient. In practice, we usually conclude the stability of an equilibrium form the solution of a D.E. (of course, this only works for simple problems). Now we give some examples of Liapunov stability of equilibrium points.

EXAMPLE 8.10.

$$\dot{x} = -\,x.$$

The exact solution is $x(t) = C\,e^{-t}$ where $C = x(0)$. Clearly $x = 0$ is an equilibrium point and it is asymptotically stable (see Fig. 8.15).

EXAMPLE 8.11. Logistic equation:

$$\dot{x} = x\,(1 - x).$$

The D.E. can be solved exactly as follows:

$$\Longrightarrow \quad \frac{dx}{x\,(1-x)} = dt$$

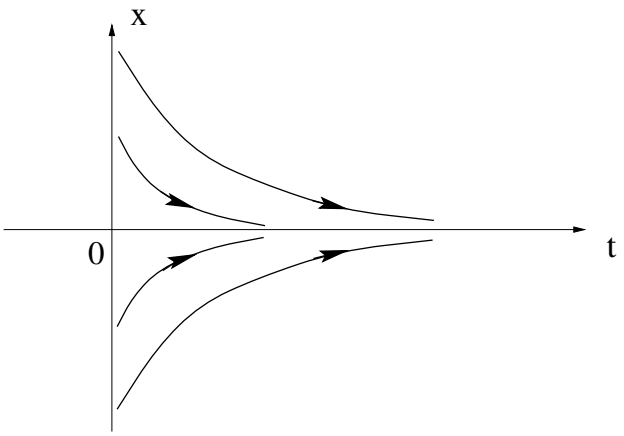


Fig. 8.15 Asymptotically stable (Example 8.10).

$$
\begin{aligned}
&\Longrightarrow \quad \int \left(\frac{1}{x} + \frac{1}{1-x}\right) dx = \int dt \\
&\Longrightarrow \quad \ln|x| - \ln|1-x| = t + \bar{C} \\
&\Longrightarrow \quad \ln\left|\frac{x}{1-x}\right| = t + \bar{C} \\
&\Longrightarrow \quad \frac{x}{1-x} = \pm\, e^{t+\bar{C}} = \pm\, e^{\bar{C}}\, e^{t} = \frac{1}{C}\, e^{t} \quad (C = \pm\, e^{-\bar{C}}) \\
&\Longrightarrow \quad x(t) = \frac{1}{1 + C\, e^{t}}\,. \qquad (8.101)
\end{aligned}
$$

If $x(0) = x_0$, then $x(t) = \dfrac{x_0}{x_0 + (1-x_0)\, e^{-t}}$. Setting $\dot{x} = 0$ yields the the equilibrium points: $x = 0$ and $x = 1$. It is easy to see that $x = 0$ is unstable while $x = 1$ is asymptotically stable. (Fig. 8.16.)

8.5.2. Stability of a Motion (Trajectory). A given motion $\tilde{\boldsymbol{y}}(t)$ is said to be stable if for any given $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that the motion $\boldsymbol{y}(t)$ following the initial disturbance for which

$$\|\boldsymbol{y}(0) - \tilde{\boldsymbol{y}}(0)\| < \delta, \qquad (8.102)$$

satisfies

$$\|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\| < \epsilon \qquad \forall \quad t > 0. \qquad (8.103)$$

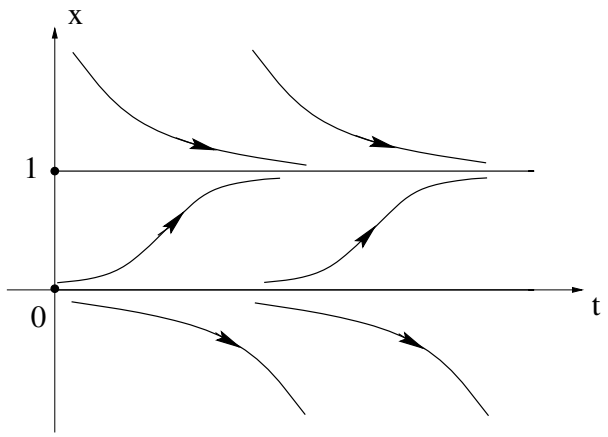


Fig. 8.16 Solutions of Example 8.11.

The motion $\tilde{\boldsymbol{y}}(t)$ is asymptotically stable if it is stable and, in addition,

$$\lim_{t\to\infty} \|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\| = 0, \tag{8.104}$$

in analogy with the case of fixed points.

In other words, a trajectory $\tilde{\boldsymbol{y}}(t)$ is stable, if all the other trajectories which start in a sufficiently close neighborhood of $\tilde{\boldsymbol{y}}(t)$ remain in a prescribed neighborhood for later times. A one dimensional case is illustrated in Fig. 8.17.

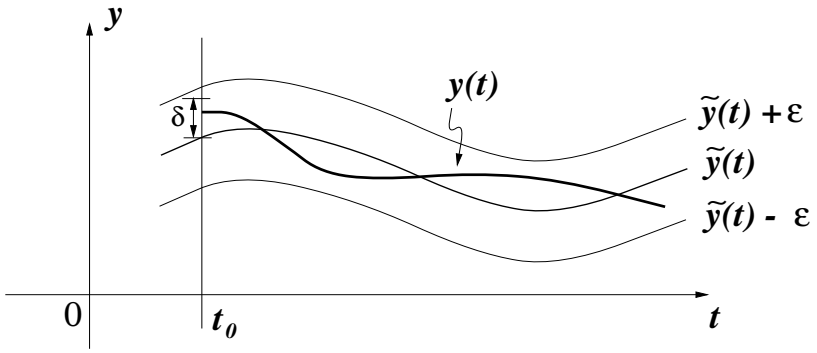


Fig. 8.17 A 1-D demonstration of stability.

It is interesting that the problem of the stability of a motion can always be reduced to the stability of the origin. Thus, in order to study the stability properties of the particular solution $\boldsymbol{y}(t) = \tilde{\boldsymbol{y}}(t)$, introduce

$$\boldsymbol{y}(t) = \tilde{\boldsymbol{y}}(t) + \boldsymbol{x}(t) \tag{8.105}$$

into equation (8.94) to obtain

$$\dot{\tilde{\boldsymbol{y}}} + \dot{\boldsymbol{x}} = \boldsymbol{f}(\tilde{\boldsymbol{y}} + \boldsymbol{x}). \tag{8.106}$$

Since $\tilde{\boldsymbol{y}}$ is a function of t, one may define

$$\boldsymbol{g}(\boldsymbol{x}, t) = \boldsymbol{f}(\tilde{\boldsymbol{y}} + \boldsymbol{x}) - \boldsymbol{f}(\tilde{\boldsymbol{y}}). \tag{8.107}$$

Observing now that

$$\dot{\tilde{\boldsymbol{y}}} = \boldsymbol{f}(\tilde{\boldsymbol{y}}) \tag{8.108}$$

leads to

$$\dot{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{y}, t) \tag{8.109}$$

where $\boldsymbol{x} = \boldsymbol{0}$ is the particular solution $\boldsymbol{y}(t) = \tilde{\boldsymbol{y}}(t)$ and $\boldsymbol{g}(\boldsymbol{0}, t) = \boldsymbol{0}$ for all t. Note, however, that the resulting system is in general non-autonomous.

There are other issues in stability which will be discussed shortly; however, we really have gone beyond stability in previous sections when we sketched phase portraits. Stability is at the heart of things, but we were also concerned with the shape and direction of the family of trajectories. We say that it is the **topology** of the flow. The *flow* is the totality of all the trajectories in some region of phase space, as if each point were some sort of fluid element.

In topology things may be stretched and distorted but still in some sense be equivalent. A thing that is topologically *equivalent* to something else is called a homeomorphism. It must have the same qualitative character. Specifically, the two *objects* in question must be related by a continuous one to one transformation (mapping).

Now we may return to a discussion of the autonomous nonlinear system: $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y})$. The following theorem describes the relation of the stabilities of the fixed point, $\boldsymbol{y}_e$, of the nonlinear DE and the fixed point, $\boldsymbol{x} = \boldsymbol{0}$ of the associated linear system $\dot{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{y}_e)\,\boldsymbol{x}$.

8.5.3. Hartman-Grobman Theorem: (Ref: Guckenheimer and Holmes, 1983, p. 13) The flow generated by $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y})$, is homeomorphic to $\dot{\boldsymbol{x}} = D\boldsymbol{f}(\boldsymbol{y}_e)\,\boldsymbol{x}$ in a neighborhood of the fixed point $\boldsymbol{y} = \boldsymbol{y}_e$, if $D\boldsymbol{f}(\boldsymbol{y}_e)\,\boldsymbol{x}$ has no zero or purely imaginary eigenvalues (i.e. hyperbolic).

This theorem together with something known as the stable manifold theorem, guarantee that one can discuss stability, and sketch phase portraits near fixed points (in any dimension of phase space) based upon the linearization for hyperbolic fixed points. If any one of the eigenvalues has zero real part, then the stability cannot be determined by linearization, leading to a critical case.

EXAMPLE 8.12.

$$\begin{aligned} \dot{x}_1 &= 2\,x_2\,(x_3 - 1) \\ \dot{x}_2 &= -\,x_1\,(x_3 - 1) \\ \dot{x}_3 &= -\,x_3^3\,. \end{aligned} \tag{8.110}$$

It is easy to find the equilibrium (only one): $(x_1\,,\,x_2\,,\,x_3) = (0,\,0,\,0)$. The derivative of (8.110) at the equilibrium point is

$$D\boldsymbol{f}(\mathbf{0}) = \begin{bmatrix} 0 & -2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

which, in turn, yields eigenvalues: $\lambda_1 = 0\,,\ \lambda_{2,3} = \pm\, i\,\sqrt{2}\,.$

The only conclusion from this example is that the origin is not a *sink*.

8.6 Stability When Linearization Does not Work: Liapunov's Direct Method

Consider

$$\ddot{x} + x + \epsilon\,x^2\,\dot{x} = 0. \tag{8.111}$$

This becomes as a system if introducing $x_1 = x$ and $x_2 = \dot{x}$:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\,x_1 - \epsilon\,x_1^2\,x_2 \end{aligned} \tag{8.112}$$

or

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \boldsymbol{f} \equiv \begin{pmatrix} x_2 \\ -\,x_1 - \epsilon\,x_1^2\,x_2 \end{pmatrix}. \tag{8.113}$$

Thus,

$$D\boldsymbol{f}(x_1,\,x_2) = \begin{bmatrix} 0 & 1 \\ -\,1 - 2\,\epsilon\,x_1\,x_2 & -\,\epsilon\,x_1^2 \end{bmatrix}, \quad \text{and} \quad D\boldsymbol{f}(0,\,0) = \begin{bmatrix} 0 & 1 \\ -\,1 & 0 \end{bmatrix}. \tag{8.114}$$

About the fixed point $(0,\,0)$ we have the linearized system

$$\dot{\boldsymbol{x}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \boldsymbol{x} = A\,\boldsymbol{x}. \tag{8.115}$$

This structure for A, we recognize as representing a matrix with purely imaginary eigenvalues. In this case, they are $\pm i$. That falls into the case where $(0,\,0)$ is a non-hyperbolic (critical) fixed point. In this case, the linear system tells us $(0,\,0)$ is a center, but the true nature will be decided by the nonlinearities which provide a phase portrait *not* homeomorphic to the linear portrait.

We can establish the stability question in this case by **Liapunov's direct method**. The idea involves some neighborhood of the fixed point, $\boldsymbol{U}$ ($\boldsymbol{U}$ is a set of points nearby the fixed point $\boldsymbol{x}_e$).

We note that $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ defines a vector field and $\boldsymbol{f}(\boldsymbol{x})$ is the rate of change of a position vector in phase space. We all know this is just tangent to the trajectory. Thus, if we image some surface or curve, $\boldsymbol{S}$, surrounding the fixed point $\boldsymbol{x}_e$ which defines $\boldsymbol{U}$, we may decide stability if we know that all of the tangent vectors are pointing inward on $\boldsymbol{S}$ or along $\boldsymbol{S}$. If it is inward or along $\boldsymbol{S}$, then we are guaranteed that once in $\boldsymbol{U}$, no trajectories can leave: thus satisfying stability. If there are a family of such curves where tangents are inward closer and closer to $\boldsymbol{x}_e$, then we know that we must arrive at $\boldsymbol{x}_e$ eventually and we have asymptotic stability. See Fig. 8.18.

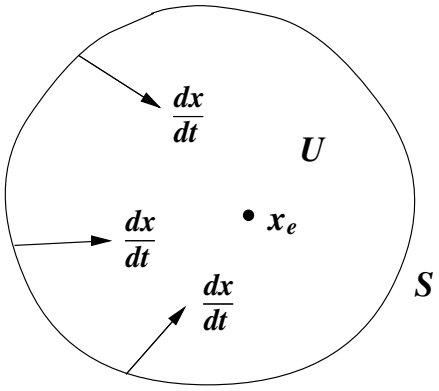


Fig. 8.18 Geometrical view of a Liapunov function.

This is fine, but how do we get an $\boldsymbol{S}$, or a family of curves $\{\boldsymbol{S}\}$?

Consider

$$W = V(x_1,\, x_2,\, \cdots\ \, x_n) = V(\boldsymbol{x}). \tag{8.116}$$

We require

$$W = 0 = V(\boldsymbol{x}_e). \tag{8.117}$$

We can either have V an increasing function or a decreasing function as one parameter $\boldsymbol{x}_e$ from all directions. It does not matter, but it is customary to make V decrease to 0 as $\boldsymbol{x} \to \boldsymbol{x}_e$. Thus, V is positive (allow it to be zero too) in $\boldsymbol{U}$; i.e.

$$V \geq 0 \qquad \text{for} \quad \boldsymbol{x} \in \boldsymbol{U}.$$

Now consider an arbitrary trajectory $\boldsymbol{x}(t) \in \boldsymbol{U}$, then

$$\frac{dV}{dt} = \bigtriangledown V \bullet \frac{d\boldsymbol{x}}{dt} \tag{8.118}$$

by the chain rule, where $\bullet$ denotes the dot product. The nature of the function (decreasing or increasing towards $\boldsymbol{x}_e$) requires $V = constant$ to make level curves or surfaces surrounding $\boldsymbol{x}_e$ essentially defining $\boldsymbol{S}$. On such a surface $\bigtriangledown V$ is perpendicular to this and pointing outward. Then

$$\frac{dV}{dt} = \bigtriangledown V \bullet \frac{d\boldsymbol{x}}{dt} \leq 0 \tag{8.119}$$

guarantees a trajectory that is not leaving the neighborhood. For arbitrary $\boldsymbol{x}(t)$, this establish stability.

Theorem: Let $\boldsymbol{x}_e$ be a fixed point of

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}).$$

Then if there exists a function $V(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{U}$, where $\boldsymbol{U}$ is a neighborhood of $\boldsymbol{x}_e$, and

i) $V(\boldsymbol{x}_e) = 0$ and $V(\boldsymbol{x}) > 0$ for $\boldsymbol{x} \neq \boldsymbol{x}_e$;

ii) $\dot{V}(\boldsymbol{x}) \leq 0$ for $\boldsymbol{x} \in \boldsymbol{U} - \{\boldsymbol{x}_e\}$;

then $\boldsymbol{x}_e$ is stable. Moreover, if

iii) $\dot{V}(\boldsymbol{x}) < 0$ for $\boldsymbol{x} \in \boldsymbol{U} - \{\boldsymbol{x}_e\}$;

then $\boldsymbol{x}_e$ is asymptotically stable. We call V a **Liapunov function**.

Now return to our example

$$\frac{d\boldsymbol{x}}{dt} = \begin{pmatrix} x_2 \\ -x_1 - \epsilon\, x_1^2\, x_2 \end{pmatrix}.$$

We note that $x_1^2 + x_2^2$ is positive for any $\boldsymbol{x} \neq \boldsymbol{0}$ and zero at $(0,\, 0)$, making a good Liapunov function candidate. Thus,

$$\begin{aligned} \dot{V} = \frac{dV}{dt} &= \nabla V \bullet \frac{d\boldsymbol{x}}{dt} \\ &= \frac{\partial V}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial V}{\partial x_2}\frac{dx_2}{dt} \\ &= 2\, x_1\, \dot{x}_1 + 2\, x_2\, \dot{x}_2 \\ &= 2\, x_1\, x_2 + 2\, x_2\, (-x_1 - \epsilon\, x_1^2\, x_2) \\ &= -2\, \epsilon\, x_1^2\, x_2^2 \end{aligned} \tag{8.120}$$

where $\dot{x}_1 = x_2$ and $\dot{x}_2 = -x_1 - \epsilon\, x_1^2\, x_2$ have been used.

The sign of $\dot{V}$ is decided by ϵ only. Thus if $\epsilon < 0 \Longrightarrow \dot{V} > 0$! All trajectories flow outward indicating that $\boldsymbol{x}_e$ is a repeller or a source. If $\epsilon > 0$, then $\dot{V} < 0$, in which case, all trajectories flow inward. Moreover, because $\dot{V} \neq 0$, except at $\boldsymbol{x}_e$, they must flow to $\boldsymbol{x}_e$: asymptotic stability. If fact, for this example, $\boldsymbol{x}_e$ is globally asymptotic stable.

The linear property of a center is topologically changed by the nonlinearity to a focus (Fig. 8.19).

Fig. 8.19 Nonlinearity changes topology:
(a) center (linear system); (b) focus (nonlinear system).

Question: How to find a Liapunov function? The following couple of examples may give you some hint.

EXAMPLE 8.13. Suppose the mass-spring system is given by

$$m\,\ddot{x} + k\,(x + x^3) = 0 \qquad (\text{"stiff spring"}). \tag{8.121}$$

Writing $\ddot{x} = \dfrac{d\dot{x}}{dt} = \dfrac{d\dot{x}}{dx}\,\dfrac{dx}{dt} = \dot{x}\,\dfrac{d\dot{x}}{dx}$ and integrating the above equation, one obtains

$$E = \frac{1}{2}\,m\,\dot{x}^2 + \frac{1}{2}\,k\,(x^2 + \frac{1}{2}\,x^4) \tag{8.122}$$

which is the total energy of the system: the first term represents the kinetic energy and the second term denotes the potential energy. Let $x_1 = x$ and $x_2 = \dot{x}$, then (8.121) becomes a 2-D system:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{k}{m}\,(x_1 + x_1^3) \end{aligned} \tag{8.123}$$

and function $E(x,\,\dot{x}) \equiv E(x_1\,,\,x_2)$. Is it a Liapunov function? Let's check the conditions:

(i) $E(0,\,0) = 0$ and $E > 0\ \ \forall\ (x_1\,,\,x_2) \neq (0,\,0,\,0)$; and

(ii) $\dot{E} = m\,x_2\,\dot{x}_2 + k\,(x_1 + x_1^3)\,\dot{x}_1 = -\,m\,x_2\,\dfrac{k}{m}(x_1 + x_1^3) + k\,(x_1 + x_1^3)\,x_2 = 0\,,$

i.e. $\dot{E} \equiv 0$; so, by the Liapunov theorem, the equilibrium $(0,\,0)$ is stable. The physical explanation of the function E is given in Fig. 8.20, where $z = E(x_1\,,\,x_2)$, and $E = \text{const.}$ represents *level* curves of E, which are closed (near ellipses). Orbits (or trajectories) are level curves of E since $\dot{E} = 0$.

EXAMPLE 8.14. Prove that the following system

$$\begin{aligned} \dot{x} &= y - x^3 \\ \dot{y} &= -\,x - y^3 \end{aligned} \tag{8.124}$$

has no closed orbits.

First, to find the fixed points of the system, letting $\dot{x} = \dot{y} = 0$ results in one real solution only, $(0,\,0)$. It can be shown that the associated linear system at $(0,\,0)$ yields a

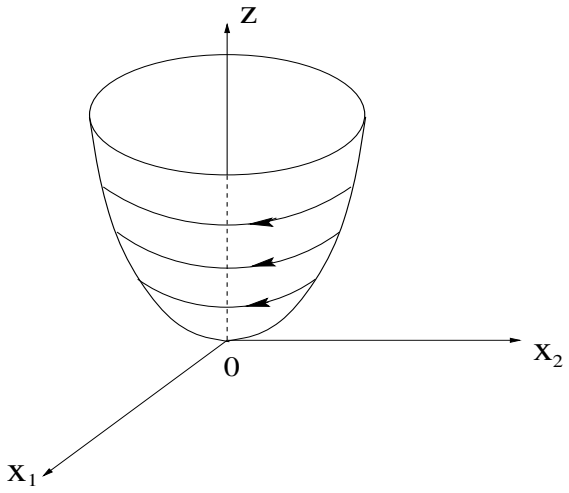


Fig. 8.20 Level curves of E (Example 8.13).

center. So, linearization method fails in this case to analyze the stability of the fixed point. Constructing a Liapunov function

$$V = a\,x^2 + b\,y^2 \qquad a,\,,\; b \;\text{ to be determined.} \tag{8.125}$$

Then

$$\begin{aligned}
\dot{V} = \frac{dV}{dt} &= \bigtriangledown V \bullet \frac{d\boldsymbol{x}}{dt} \\
&= \frac{\partial V}{\partial x}\,\frac{dx}{dt} + \frac{\partial V}{\partial y}\,\frac{dy}{dt} \\
&= 2\,a\,x\,\dot{x} + 2\,b\,y\,\dot{y} \\
&= 2\,a\,x\,(y - x^3) + 2\,b\,y\,(-\,x - y^3) \\
&= 2\,(a-b)\,x\,y - 2\,(a\,x^4 + b\,y^4)\,.
\end{aligned} \tag{8.126}$$

If choose $a = b > 0$, then

$$\dot{V} = -2\,a\,(x^4 + y^4) < 0 \qquad \forall \quad (x,\,y) \neq (0,\,0). \tag{8.127}$$

Therefore, the point $(0,\,0)$ is globally asymptotically stable. In other words, all trajectories (no matter starting from what initial point) converge to the origin because the Liapunov function is valid for whole space $\boldsymbol{R}^2$. So no closed orbits can exist.

Chapter 9

CYCLES: FLOQUET THEORY

So far we have discussed stability about fixed points but we can, as should be clear in our linearization discussion in Section 8.1, discuss stability about solutions in general. In this case, our linear system has the form

$$\frac{d\boldsymbol{x}}{dt} = D\boldsymbol{f}(\tilde{\boldsymbol{y}})\,\boldsymbol{x} \tag{9.1}$$

where $\tilde{\boldsymbol{y}}$ is our reference solution. Suppose now that $\tilde{\boldsymbol{y}}$ maps out a closed orbit with period T (i.e. $\tilde{\boldsymbol{y}}$ is a periodic function). Then

$$D\boldsymbol{f}(\tilde{\boldsymbol{y}}) \equiv P(t) = P(t+T). \tag{9.2}$$

So we consider the differential equation

$$\frac{d\boldsymbol{x}}{dt} = P(t)\,\boldsymbol{x} \tag{9.3}$$

where every element of $P(t)$ is a periodic function.

Remark: In general, this kind of DEs can be written as an n-th order

$$\frac{d^n x}{dt} + a_1(t)\,\frac{d^{n-1}x}{dt} + \ldots + a_{n-1}(t)\,\frac{dx}{dt} + a_n(t)\,x = 0$$

where all $a_i(t)\,,\ i = 1,\, 2,\, \ldots\, n$ are periodic functions, or in a system of equations:

$$\frac{d\boldsymbol{x}}{dt} = A(t)\,\boldsymbol{x} \tag{9.4}$$

where $A(t)$ is an $n \times n$ matrix whose elements are periodic functions. The solutions of the DE are, in general, similar to that of a DE with constant coefficients. It is one important type of DEs which arises frequently in mechanics and physics. However, it should be pointed out that although we can generally discuss the solutions of the DE (e.g. stability), we do

not have any power methods to find the solutions. Numerical computations may still be the best way to solve this kind of DEs.

We note that from the above equation that

$$\frac{d\boldsymbol{x}(t+T)}{dt} = P(t+T)\,\boldsymbol{x}(t+T) \tag{9.5}$$

which implies that

$$\frac{d\boldsymbol{x}(t+T)}{dt} = P(t)\,\boldsymbol{x}(t+T) \tag{9.6}$$

or $\boldsymbol{x}(t+T)$ is a solution too, but with different initial conditions than $\boldsymbol{x}(t)$. That is, it is a particular solution of the DE. If we use $\Phi(t)$ as a matrix consisting of n linearly independent solutions, which is called basic solution matrix. (The Wronskian $W = \det(\Phi(t)) \neq 0$ for any t.) Then

$$\frac{d\boldsymbol{x}(t+T)}{dt} = \Phi(t)\,\boldsymbol{a} \tag{9.7}$$

and

$$\frac{d\boldsymbol{x}(t)}{dt} = \Phi(t)\,\boldsymbol{b} \tag{9.8}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant column vectors representing the different initial conditions.

It follows since $\boldsymbol{a}$ and $\boldsymbol{b}$ are just constants that

$$\frac{d\Phi(t)}{dt} = P(t)\,\Phi(t) \qquad \text{where} \qquad \Phi(0) = I \tag{9.9}$$

which indicates that $\Phi(t)$ is a solution, and $\Phi(t+T)$ is a solution too. Consequently

$$\Phi(t+T) = \Phi(t)\,M \tag{9.10}$$

where M is a constant matrix. Thus, in general, after m periods (m is an arbitrary natural number)

$$\Phi(t) = \Phi(t'+mT) = \Phi(t')\,M^m = \Phi(t')\,e^{TBm} \tag{9.11}$$

where $e^{TBm} \equiv M^m$, which implies $e^{TB} = M$, and so B is a constant matrix. Therefore (note that $m = \dfrac{t-t'}{T}$),

$$\Phi(t) = \Phi(t')\,e^{(t-t')\,B} = \Phi(t')\,e^{-Bt'}\,e^{Bt} = Q(t)\,e^{Bt} \tag{9.12}$$

where the property of commutation of $e^{-Bt'}$ with e^{Bt} has been used. Here, $Q(t) \equiv \Phi(t')\, e^{-Bt'}$ is a periodic function by definition since it depends upon t through m only $t' = t - mT$. And so,

$$\Phi(t)\, \boldsymbol{b} = Q(t)\, e^{Bt}\, \boldsymbol{b} = \boldsymbol{x}(t). \tag{9.13}$$

Thus the asymptotic behavior of $\boldsymbol{x}(t)$ depends upon the eigenvalues of B.

Now, let's try to find the nonzero solutions of the DE

$$\frac{d\boldsymbol{x}(t)}{dt} = P(t)\, \boldsymbol{x}(t) \tag{9.14}$$

satisfying

$$\boldsymbol{x}(t+T) = \lambda\, \boldsymbol{x}(t), \tag{9.15}$$

where λ is a constant to be determined. We call such a solution a regular solution of the DE. It is easy to see that when $\lambda = 1$, the DE has periodic solutions with period T; and when $\lambda = -1$, it has periodic solutions with period $2T$. We want to find the equation that the λ must satisfy. First, we can express an arbitrary nonzero solution $\boldsymbol{x}(t)$ as

$$\boldsymbol{x}(t) = \Phi(t)\, \boldsymbol{c} \tag{9.16}$$

where $\boldsymbol{c} \neq \boldsymbol{0}$. Then

$$\boldsymbol{x}(t+T) = \Phi(t+T)\, \boldsymbol{c} = \Phi(t+T)\, M\, \boldsymbol{c} \tag{9.17}$$

where the relation $\Phi(t+T) = \Phi(t)\, M$ has been used. Therefore, substitute the above two equations into equation (9.15) to obtain

$$\Phi(t)\ M\, \boldsymbol{c} = \lambda\, \Phi(t)\, \boldsymbol{c} = \Phi(t)\, \lambda\, \boldsymbol{c}, \tag{9.18}$$

that is,

$$\Phi(t)\, (\lambda\, I - M)\, \boldsymbol{c} = \boldsymbol{0}. \tag{9.19}$$

Since $\det(\Phi(t)) \neq 0$ and $\boldsymbol{c} \neq \boldsymbol{0}$, the above equation implies that

$$\det\, (\lambda\, I - M) = 0 \tag{9.20}$$

which is actually the characteristic polynomial of the matrix M. The roots of the polynomial are the eigenvalues of M. This polynomial is invariant under different basic solution matrix

$\Phi(t)$. It should be noted, however, that this case is different from that the DE's with constant coefficients. For the latter case, the characteristic polynomial can be found directly from DEs, while for the former case a basic solution matrix must be found first in order to find matrix M from the equation $\Phi(t+T) = \Phi(t)\, M$. Thus, here we cannot obtain solutions based on the characteristic polynomial, like we did in solving DEs with constant coefficients. However, we can still discuss the structure and properties of the solutions.

Suppose the eigenvalues of B be σ_i , $i = 1, 2, \ldots n$. We can establish the relation between σ_i and λ_i as follows. First, assume that the eigenvalues λ_i of M are distinct, then we can find a non-singular matrix S such that

$$S^{-1}MS = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} = \begin{bmatrix} e^{\ln\lambda_1} & 0 & \ldots & 0 \\ 0 & e^{\ln\lambda_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & e^{\ln\lambda_n} \end{bmatrix} = \exp \begin{bmatrix} \ln\lambda_1 & 0 & \ldots & 0 \\ 0 & \ln\lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ln\lambda_n \end{bmatrix}.$$

and so

$$e^{S^{-1}TBS} = e^{TS^{-1}BS} = S^{-1}\, e^{TB}\, S = S^{-1}\, M\, S = \exp \begin{bmatrix} \ln\lambda_1 & 0 & \ldots & 0 \\ 0 & \ln\lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ln\lambda_n \end{bmatrix} \tag{9.21}$$

which gives

$$S^{-1}BS = \begin{bmatrix} \frac{1}{T}\ln\lambda_1 & 0 & \ldots & 0 \\ 0 & \frac{1}{T}\ln\lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{T}\ln\lambda_n \end{bmatrix}. \tag{9.22}$$

From the above equation, we can write the characteristic polynomial of S as

$$\begin{aligned} 0 &= \det(\sigma\, I - S) = \det(S)\det(\sigma\, I - S^{-1}\, B\, S)\det(S^{-1}) = \det(\sigma\, I - S^{-1}\, B\, S) \\ &= \det \begin{bmatrix} \sigma - \frac{1}{T}\ln\lambda_1 & 0 & \ldots & 0 \\ 0 & \sigma - \frac{1}{T}\ln\lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma - \frac{1}{T}\ln\lambda_n \end{bmatrix} = \Pi_{i=1}^{n}\,(\sigma - \frac{1}{T}\ln\lambda_i) \end{aligned} \tag{9.23}$$

which, in turn, yields

$$\sigma_i = \frac{1}{T} \ln \lambda_i \,, \quad i = 1, 2, \ldots n. \tag{9.24}$$

The σ_i's are called **characteristic multipliers** of the DE. Thus, each element of the solution $\boldsymbol{x} = Q(t)\, e^{Bt}\, \boldsymbol{b}$ is now in the form

$$x_i = c_{i1}\, \phi_{i1}(t)\, e^{\sigma_1 t} + c_{i2}\, \phi_{i2}(t)\, e^{\sigma_2 t} + c_{i3}\, \phi_{i3}(t)\, e^{\sigma_3 t} + \ldots + c_{in}\, \phi_{in}(t)\, e^{\sigma_n t} \tag{9.25}$$

where $\phi_{ij}(t)$ are all periodic functions. It is in the similar form of the solutions of DE's with constant coefficients (see Section 8.2). Therefore, we have the following conclusions:

1. If all the multipliers, σ_i, of the DE have negative real parts (i.e. all the eigenvalues λ_i of M have modulus less than 1), then $\boldsymbol{x}(t) \to \boldsymbol{0}$ as $t \to \infty$.

2. If at least one of the multipliers, σ_i, have positive real parts (i.e. the corresponding eigenvalues λ_i of M have modulus great than 1), then there exist solutions $\boldsymbol{x}(t)$ unbounded for $t \in [0, \infty)$.

3. If some or all of the multipliers, σ_i, have zero real parts (i.e. the corresponding eigenvalues λ_i of M have modulus 1), and the rest have negative real parts, then every solution $\boldsymbol{x}(t)$ is bounded for $t \in [0, \infty)$.

For the repeated roots, a similar discussion as that given in Section 8.2 for the DE with constant coefficients can be carried out here. For example, suppose $\sigma_1 = \sigma_2$, then every element of the solution $\boldsymbol{x}$ can be expressed as

$$x_i = c_{i1}\, \phi_{i1}(t)\, e^{\sigma_1 t} + c_{i2}\, \phi_{i2}(t)\, t\, e^{\sigma_1 t} + c_{i3}\, \phi_{i3}(t)\, e^{\sigma_3 t} + \ldots + c_{in}\, \phi_{in}(t)\, e^{\sigma_n t}. \tag{9.26}$$

Thus, in the case of repeated roots, we have

1) If all the multipliers, σ_i, have negative real parts, then $\boldsymbol{x}(t) \to \boldsymbol{0}$ as $t \to \infty$.

2) If at least one of the multipliers, σ_i, have zero or positive real parts, then there exist solutions $\boldsymbol{x}(t)$ unbounded for $t \in [0, \infty)$.

EXAMPLE 8.1. Discuss the solution of the following DE

$$\frac{dx}{dt} = a(t)\, x, \tag{9.27}$$

where $a(t)$ is a periodic function with period T.

Solution: The solution of the DE is

$$x = x(t) \equiv x(0)\, e^{\int_0^t a(\tau)\, d\tau}, \tag{9.28}$$

so,

$$x(t+T) \equiv x(0)\, e^{\int_0^{t+T} a(\tau)\, d\tau}. \tag{9.29}$$

But,

$$\int_0^{t+T} a(\tau)\, d\tau \equiv \int_0^t a(\tau)\, d\tau + \int_t^{t+T} a(\tau) d\tau \equiv \int_0^t a(\tau)\, d\tau + \int_0^T a(\tau)\, d\tau, \tag{9.30}$$

therefore,

$$x(t+T) \equiv x(0)\, e^{\int_0^t a(\tau)\, d\tau}\, e^{\int_0^T a(\tau)\, d\tau} \equiv x(t)\, e^{\int_0^T a(\tau)\, d\tau}. \tag{9.31}$$

The above equation indicates that the matrix (which is a scaler for this example) $M = e^{\int_0^T a(\tau)\, d\tau}$, and so the characteristic polynomial is given by

$$\lambda - e^{\int_0^T a(\tau)\, d\tau} = 0. \tag{9.32}$$

Hence, the eigenvalue of the DE is $\lambda = e^{\int_0^T a(\tau)\, d\tau}$, and the multiplier $\sigma = \dfrac{1}{T} \displaystyle\int_0^T a(\tau)\, d\tau$. The solution of the DE can now be written as

$$x(t) \equiv x(0)\, e^{\int_0^t a(\tau)\, d\tau - \sigma t}\, e^{\sigma t} \equiv \phi(t)\, e^{\sigma t}, \tag{9.33}$$

where $\phi(t) \equiv x(0)\, e^{\int_0^t a(\tau)\, d\tau - \sigma t}$ is a periodic function with period T. It can be seen that the DE has periodic solutions with period T only when $\sigma = \dfrac{1}{T} \displaystyle\int_0^T a(\tau)\, d\tau = 0$.

Chapter 10

ASYMPTOTIC BEHAVIOR: ATTRACTORS AND REPELLERS

We have already mentioned that a fixed point cannot be arrived in finite time, simply because of uniqueness. This must also be true of cyclical behavior. Fig. 10.1 shows that the "damping" of some trajectory to a closed cycle, as suggested by the Floquet theory in the previous chapter, excludes arriving in finite time at the cycle. (We can also have unstable cycles – of course – depending on the matrix B given in the previous chapter.)

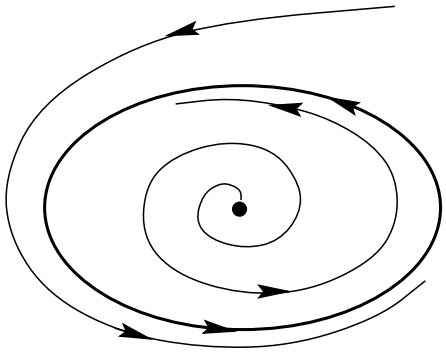


Fig. 10.1 Phase portrait of a stable cycle.

This behavior must be asymptotic in any case. However for practical purpose, we have already seen that this asymptotic behavior is practically achieved.

Thus, we may expect, as we did with maps, that we may have periodic behavior resulting in nonlinear systems (having nothing to do with the usual trigonometric functions) as a result of some relaxation to a closed orbit in phase space. This cycle is referred to as a limit cycle. There is a whole field connected with nonlinear waves. Nonlinear oscillation is another filed exhibiting such phenomena. Note that even for an autonomous system, one may find limit cycles.

EXAMPLE 10.1. Consider the following 2-dimensional system:

$$\begin{aligned} \dot{x} &= -y + a\,x\,(1 - x^2 - y^2) \\ \dot{y} &= \quad x + a\,y\,(1 - x^2 - y^2) \end{aligned} \tag{10.1}$$

where the constant $a > 0$. To find the solutions of the system, use the transformation

$$\begin{aligned} x &= r\,\cos\theta \\ y &= r\,\sin\theta \end{aligned} \tag{10.2}$$

and differentiate it, and then substitute them into the DE (10.1) to obtain

$$\begin{aligned} \dot{r}\,\cos\theta - r\,\dot{\theta}\,\sin\theta &= \dot{x} = -r\,\sin\theta + a\,r\,(1 - r^2)\,\cos\theta \\ \dot{r}\,\sin\theta + r\,\dot{\theta}\,\sin\theta &= \dot{x} = \quad r\,\cos\theta + a\,r\,(1 - r^2)\,\sin\theta. \end{aligned} \tag{10.3}$$

Multiplying the 1st equation of (10.3) by $\cos\theta$ and the 2nd equation by $\sin\theta$, and adding the resulting equations result in

$$\dot{r} = a\,r\,(1 - r^2). \tag{10.4}$$

Similarly, one can obtain

$$r\,\dot{\theta} = r \qquad \text{i.e.} \qquad \dot{\theta} = 1. \tag{10.5}$$

The solution of the above equations can be easily obtained as

$$\begin{aligned} r(t) &= \frac{1}{\sqrt{1 - (1 - r_0^{-2})\,e^{-2\,a\,t}}}, \\ \theta(t) &= t + \theta_0\,, \end{aligned} \tag{10.6}$$

where r_0 and θ_0 are initial conditions. It is seen that if $r_0 = 1$, then $r(t) \equiv 1$. If $r_0 \neq 1$, then $r(t) \to 1$ as $t \to \infty$ due to $a > 0$. It is a stable limit cycle.

For asymptotic behavior we are usually most interested in bounded behavior. We do want to characterize in some sense the ultimate outcome of starting along some trajectory: going to infinity is a possibility but uninteresting possibility.

We define limit points, but simply asking where one ends up doesn't work if the trajectory ends in a cycle. We want to take into account the possibility that the "final outcome" is not a fixed point. We also want some scope, regarding what a whole region of phase space does.

ω limit point: $\boldsymbol{x}_0 \in \boldsymbol{R}^n$ is an ω limit point of $\boldsymbol{x} \in \boldsymbol{R}^n$ if there exists a sequence in $t,\ t_0\,,\ t_1\,,\ \ldots$ as $t_i \to \infty$ such that

$$\phi(t_i,\ \boldsymbol{x}) \to \boldsymbol{x}_0 \tag{10.7}$$

where $\phi(t_i,\ \boldsymbol{x})$ represents a family of trajectories, each member denoted by the reference starting points, $\boldsymbol{x}$. (See Fig. 10.2.) That is, ϕ is the *flow*. Thus as a trajectory relax onto a limit cycle a point on the cycle can be an ω limit.

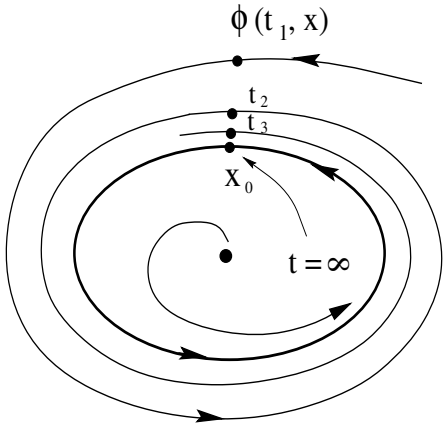


Fig. 10.2 An example of ω limit sets.

If we let "t" vary continuously there is, of course, no limit point. An α **limit point** is defined as $t_i \to -\infty$. The set of all ω limit points of flow is the ω limit set. Thus, the limit cycle is the ω limit set of the surrounding flow.

A saddle point, shown in Fig. 10.3, is both an ω and α limit set, depending upon trajectories.

An **attracting set** is an invariant set that attracts all surrounding trajectories. A stable fixed point is the simplest attractor. The set of points formed by a stable limit cycle is an attractor. A repelling set "repels" surrounding trajectories. A saddle is neither. The ω limit set of the limit cycle is an attractor.

Basins of an attractor are those regions of phase space attracted to a particular attractor.

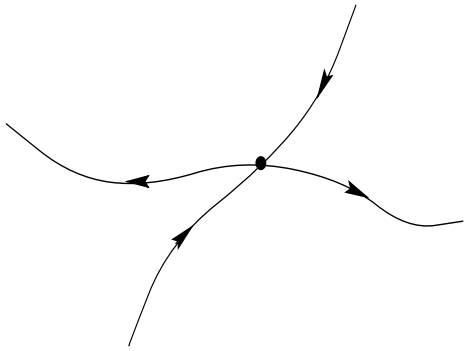


Fig. 10.3 A saddle point.

EXAMPLE 10.2. The damped Duffing oscillator described by the following equation

$$\ddot{x} + \alpha\,\dot{x} - x + x^3 = 0 \qquad \alpha > 0 \tag{10.8}$$

is an example of attracting basin (Fig. 10.4)

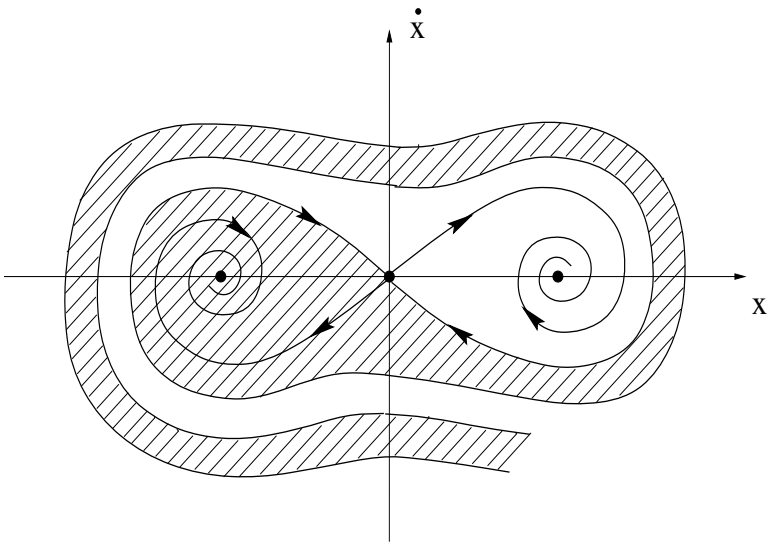


Fig. 10.4 An example of attracting basin.

10.1 Index Theory: Restrictions in the Phase Plane

It should be clear that fixed points and orbits represent the key properties of a phase portrait in the plane (we formalize this in the next section). The fact remains that these

"building blocks" cannot be put together freely. There are restrictions that will tell us a lot about how a particular DE can behave. Index theory is a simple example.

In the plane the autonomous system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1,\, x_2) \\ f_2(x_1,\, x_2) \end{pmatrix} \tag{10.9}$$

represents a vector field (as already stated). In particular $\vec{x}$ is tangent to trajectories. Existence also ensures that it represents a vector at every point (except fixed points, of course) in the plane.

We have learned how to linearize a system of differential equations about a fixed point and discussed the phase portrait near the fixed point. Linearization is a method for analyzing the local behavior of the system. It does not tell us what happens to the trajectories after they leave the neighborhood of the fixed point. Furthermore, if the vector field starts with quadratic on higher-order terms, the linearization tells us nothing.

The index theory can provide *global* information about the phase portrait. It is mainly concerned with two questions: must a closed trajectory always encircle a fixed point? If so, what types of fixed points are permitted? The method also yields information about the trajectories near higher-order fixed points. Finally, we can sometimes use index arguments to rule out the possibility of closed orbits in certain parts of the phase plane.

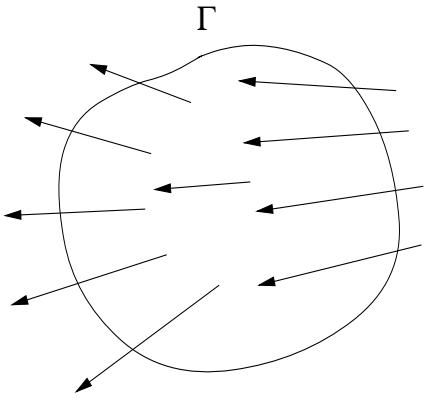


Fig. 10.5 Closed loop Γ.

Suppose we draw a closed loop Γ, which is *not* necessarily a trajectory (or a orbit, or a solution curve) in the plane, and we move counterclockwise around Γ (Fig. 10.5). As we

move around Γ, $\boldsymbol{x}$ may have different directions. The **index** of a closed curve Γ is defined as an integer that measures the winding of the vector field on Γ. If Γ surrounds no sort of asymptotic objects (cycles or fixed points), then the angle that $\boldsymbol{x}$ makes with the x_1 axis will remain unchanged after traveling around Γ once. We say the index is "0" for no change, and denote it as $I = 0$.

If curve about is a source, sink, center, or a closed orbit of the flow, then the angle will be 2π: *one* whole rotation. (See Fig. 10.6 for the case of source.) Then we say the index is $+1$, i.e. $I = +1$. This shows that the index is not related to *stability*.

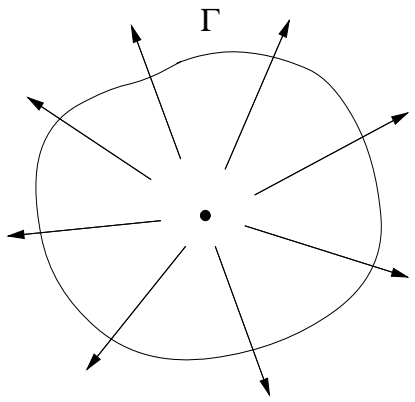


Fig. 10.6 Index for a source.

For a saddle the index $I = -1$ (i.e. -2π rotation, see Fig. 10.7). This indicates that a saddle point is truly a different animal from all the other familiar types of isolated fixed points: sink, source, center, etc.

The final point is the relation between the index of a closed curve and the indices of the fixed points inside it, which is stated as the following theorem.

Index Theorem. If a closed curve Γ surrounds n isolated fixed points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ... $\boldsymbol{x}_n$, then

$$I_\Gamma = I_1 + I_2 + \ldots I_n \tag{10.10}$$

where I_k is the index of the fixed point $\boldsymbol{x}_k$ for $k = 1, 2, \ldots n$.

Thus, any closed **orbit** (trajectory, not an arbitrary closed curve) in the phase plane must enclose fixed points whose indices sum to $+1$. This implies

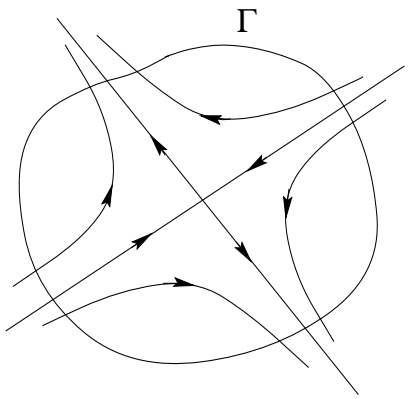


Fig. 10.7 Index for a saddle.

1. that inside any closed orbit there must be at least one fixed point.
2. If there is one, it must be a source, sink, or center (not a saddle).
3. If the fixed points are hyperbolic there must be an odd number: n saddles and $n+1$ sinks or sources.

For example, the diagrams shown in Fig. 10.8 cannot depict phase space structure for differential equations.

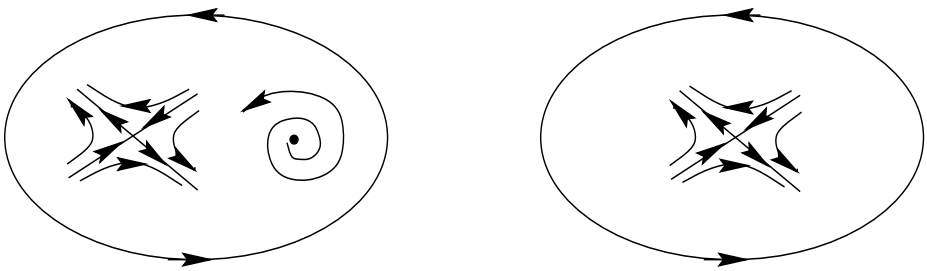


Fig. 10.8 Not existed phase portraits in 2-D system.

PROOF OF INDEX THEOREM: The argument is a familiar one, and comes up in multivariable calculus, complex variables, electrostatics, and various other subjects. We deform the curve C in such a way that the new closed curve Γ consists of n small circles $\gamma_1, \gamma, \cdots, \gamma_n$ about the fixed points, and two-way bridges connecting these circles (Fig. 10.9). Note that $I_\Gamma = I_C$ (by the property of the index: if C can be continuously deformed into C' without

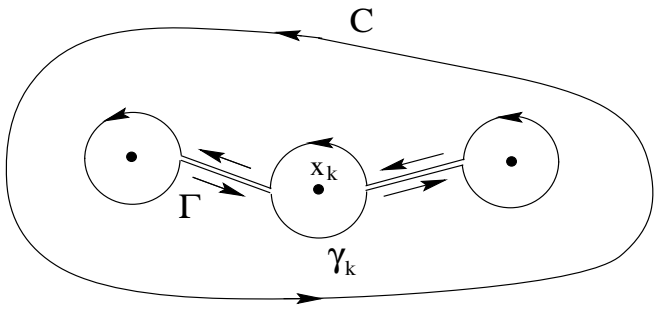


Fig. 10.9 Proof of the index theory.

passing through a fixed point. Then $I_C = I_{C'}$), since we did not cross any fixed points during the deformation. Now let's compute I_Γ by considering $[\phi]_\Gamma$. There are contributions to $[\phi]_\Gamma$ from the small circles and from the two-way bridges. However, the contributions from the two-bridges cancel out: as we move around Γ, each bridge is traversed once in one direction, and later in the opposite direction. Thus we only need to consider the contributions from the small circles. On Γ_k, the angle ϕ changes by $[\phi]_{\Gamma_k} = 2\,\pi\,I_k$, by definition of I_k. Hence

$$I_\Gamma = \frac{1}{2\,\pi}\,[\phi]_\Gamma = \frac{1}{2\,\pi}\sum_{k=1}^{n}[\phi]_{\Gamma_k} = \sum_{k=1}^{n} I_k \tag{10.11}$$

and since $I_\Gamma = I_C$, the theorem is proved.

EXAMPLE 10.3. A closed orbit in the phase plane encircles S saddles, N nodes, F spirals, and C centers, all of the usual type. Show that $N + F + C = 1 + S$.

Solution: Let the closed orbit be Γ, then $I_\Gamma = 1$. We also know that $I_S = -\,S$, $I_N = +\,N$, $I_F = +\,F$ and $I_C = +\,C$. Thus, the index theory implies that $I_\Gamma = I_S + I_N + I_F + I_C$, i.e. $1 = -\,S + N + F + C$, so, $N + F + C = 1 + S$.

EXAMPLE 10.4. Use index theory to show that the system

$$\begin{aligned} \dot{x} &= x\,(4 - y - x^2) \\ \dot{y} &= y\,(x - 1) \end{aligned} \tag{10.12}$$

has no closed orbits.

Solution: First, find all possible fixed points which are obtained by setting $\dot{x} = \dot{y} = 0$: $(0,\,0)$, $(2,\,0)$, $(-\,2,\,0)$ and $(1,\,3)$. Then it is easy to use the linearization about these fixed

points to classify them as follows. $(0,\ 0)$ and $(2,\ 0)$ are saddle points, while $(-2,\ 0)$ is a (stable) sink (node) and $(1,\ 3)$ is a (stable) focus (spiral). So, their indices are -1 -1, $+1$ and $+1$, respectively. Note that x and y coordinates are invariant (i.e. if a trajectory starting on x or y axis, it will stay on it forever). (See Fig. 10.10.) Therefore, a closed orbit cannot cross the x and y axes due to uniqueness. Further, a closed orbit must enclose fixed points whose indices sum to $+1$. Therefore, the only possibilities of closed orbits which would have to encircle the sink or the focus. But such an orbit can't encircle the sink because, otherwise, the orbit would cross the x-axis. Similarly, orbit can't encircle the focus, since the focus is joined to saddle at $(2,\ 0)$ by a branch of the saddle's unstable manifold, and orbit can't cross this trajectory.

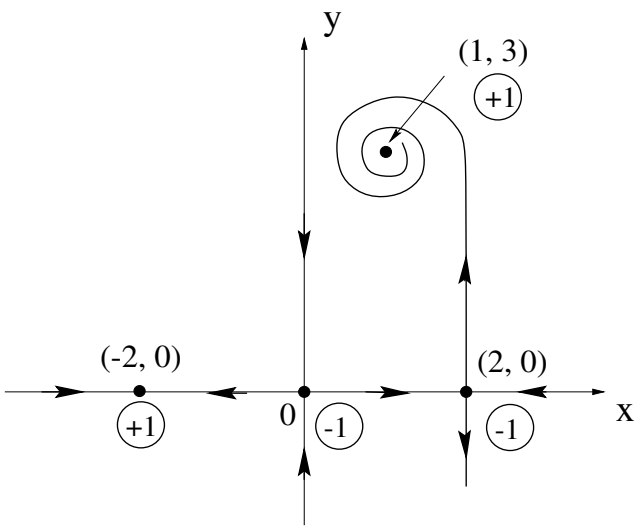


Fig. 10.10 Indices for Example 10.4.

10.2 Poincaré-Bendixson Theorem

The index theory we learned in the last section can be used sometimes to rule out closed orbits. Now we introduce the **Poincaré-Bendixson Theorem** which can be applied to establish that closed orbits exist in particular systems. Although the Poincaré-Bendixson Theorem applies in the phase plane, an important part of its significance is what it says about chaos: that is, chaos cannot occur in the plane.

Poincaré-Bendixson Theorem: Let Ω be a region where the flow is happened (doesn't wander around). Let $P \in \Omega$ and consider the ω limit set of P. Then the following possibilities hold for ω limit points.

(i) $\omega(P)$ is a fixed point.

(ii) $\omega(P)$ is a closed orbit.

(iii) $\omega(P)$ consists of a finite number of fixed points joined by individual curves (separators).

Note in the latter case, that there can be only one curve pointing two distinct fixed points in the same time sense. For example, an ω limit set, shown in Fig. 10.11, has only one ruled out portrait like Fig. 10.12.

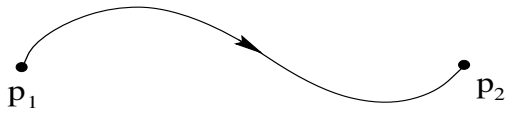


Fig. 10.11 Distinct fixed points.

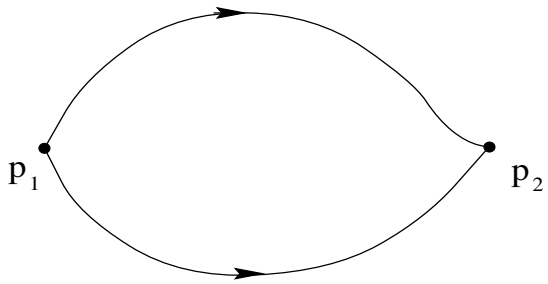


Fig. 10.12 Impossible trajectories.

However, something like shown in Fig. 10.13 can exist.
Fixed points and curves are clearly not chaotic limit set! Therefore, no chaos can exist in 2-dimensional phase space.

If one can construct a region R such that

(1) R is a closed, bounded subset of the plane;

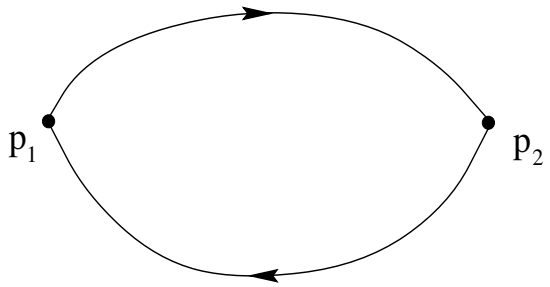


Fig. 10.13 Some particular trajectories.

(2) R does not contain any fixed points; and

(3) The continuous differential vector field $\dfrac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x})$ points *inward* everywhere on the boundary of R.

Then the Poincaré-Bendixson theorem guarantees that R contains a closed orbit. Such a region R is called *trapping region*. See Fig. 10.14.

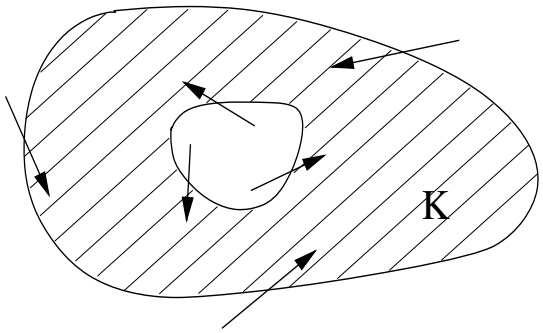


Fig. 10.14 Trapping region R.

The idea behind the Poincaré-Bendixson theorem is that the trajectories cannot wander around forever; they will finally cross if do not converge to some closed curve. However, in practice, the difficulty is how to construct such a R. The following example illustrates how to use the the Poincaré-Bendixson theorem.

EXAMPLE 10.5. Use Poincaré – Bendixson theorem to prove the existence of a periodic

attractor in the system described by

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} x - y - x\,(x^2+y^2) \\ x + y - y\,(x^2+y^2) \end{pmatrix} \equiv \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{10.13}$$

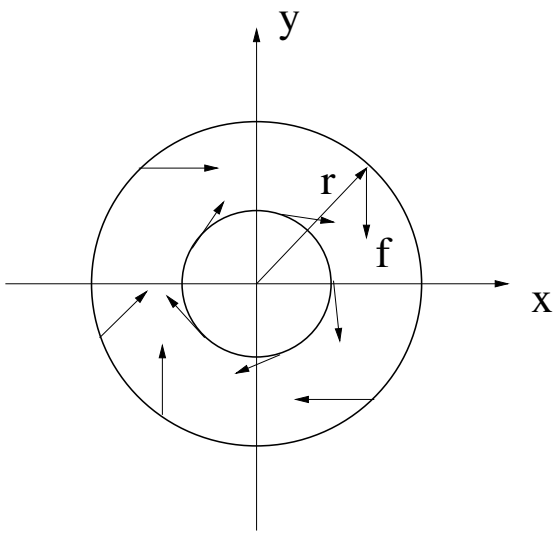


Fig. 10.15 The trapping region R for Example 10.5.

We want to find a region R such that the vector filed $\boldsymbol{f}$ points inward on the boundary of R. Consider a circle of radius D, and let vector $\boldsymbol{r} = x\,\boldsymbol{i} + y\,\boldsymbol{j}$ being a point on the circle. Computing the inner product of $\boldsymbol{r}$ and $\boldsymbol{f}$ yields

$$\begin{aligned} < \boldsymbol{r}\,,\ \boldsymbol{f} > &= x\left[x - y - x\,(x^2+y^2)\right] + y\left[x + y - y\,(x^2+y^2)\right] \\ &= x^2 + y^2 - (x^2+y^2)^2 \\ &= (x^2+y^2)\,(1 - x^2 - y^2) \\ &= D^2\,(1 - D^2) \\ &= \begin{cases} > 0 & \text{for } D^2 < 1 \\ < 0 & \text{for } D^2 > 1\,. \end{cases} \end{aligned} \tag{10.14}$$

So, let $R =$ annulus defined by

$$R = \{(x,\,y) \mid D_1 \le \sqrt{x^2+y^2} \le D_2\,,\ \ D_1 < 1 < D_2\}\,, \tag{10.15}$$

then $\boldsymbol{f}$ points to interior of R on its boundary. (See Fig. 4.10.) Therefore, we can apply Poincaré – Bendixson theorem to conclude that R contains a closed orbit. In fact, if we choose $D_1 \to 1^-$ and $D_2 \to 1^+$ we can find that the closed orbit is the unit circle.

Chapter 11

INTRODUCTION TO CHAOS

11.1 Introduction

For some people, the study of dynamics began and ended with Newton's Law $F = m\,a$. Many of us might thought that if the forces between particles and their initial positions and velocities were given, we could predict the motion or history of a system forever into the future, given a big enough computer. However, the arrival of large and fast computers has not fulfilled the promise of infinite predictability in dynamics. We now know that the motion of very simple dynamical systems cannot always be predicted far into the future. Such motions have been called **chaotic**, and their study has promoted a discussion of some exciting new mathematical ideas in dynamics. Three centuries after the publication of Newton's *Principia* (1687), it is appropriate that new phenomena have been discovered in dynamics and that new mathematical concepts from topology and geometry have entered this venerable science.

In the physical science, the paragon of chaotic phenomena is turbulence. Thus, a rising column of smoke or the eddies behind a boat or aircraft wing provide graphic examples of chaotic motion. A great deal of the excitement in nonlinear dynamics today is centered around the hope that the transition from ordered to disordered flow may be explained or modeled with relatively simple mathematical equations. It is recognition that chaotic dynamics are inherent in all of nonlinear physical phenomena that has created a sense of revolution in science and engineering today.

We must distinguish here between so-called random and chaotic motions. The former is reserved for problems in which we truly do not know the input forces or we only know some statistical measures of the parameters. the term *chaotic* is reserved for those *deterministic* problems for which there are no random or unpredictable inputs or parameters. The existence of chaotic or unpredictable motions from the classical equations of physics was known by Poincaré. Consider the following excerpt from his essay on *Science and Method*:

It may happen that small difference in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible.

In the current literature, *chaotic* is a term assigned to that class of motions in deterministic physical and mathematical systems whose time history has a *sensitive dependence on initial conditions.*

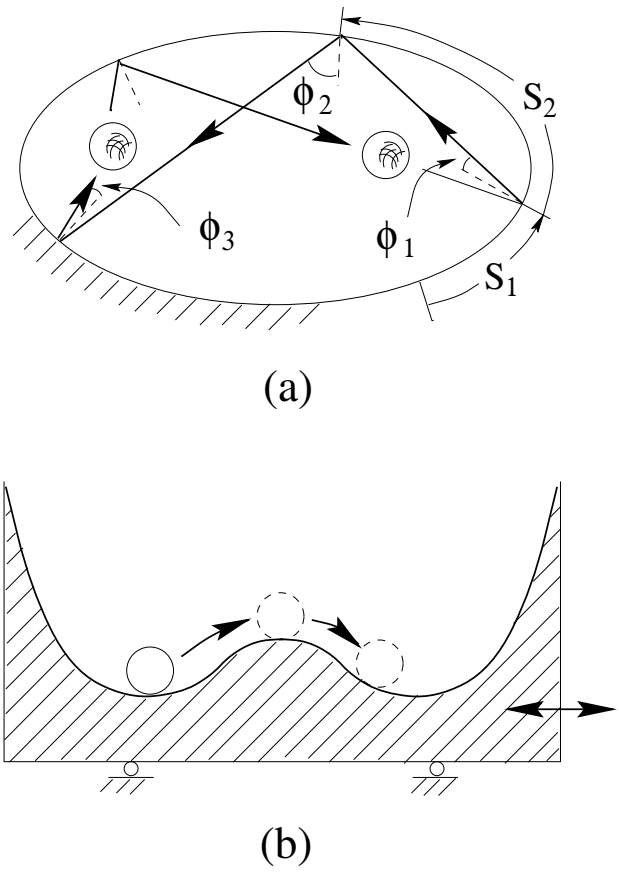


Fig. 11.1 Examples of chaos: (a) a ball with an elliptically shaped table; (b) a particle in a two-well potential.

Two examples of mechanical systems that exhibit chaotic dynamics are shown in Fig. 11.1. Fig. 11.1(a) shows a thought experiment of an idealized billiard ball (rigid body rotation is neglected) which bounces off the sides of an elliptically shaped billiard table. When elastic impact is assumed, the energy remains conserved, but the ball may wander

around the table without exactly repeating a previous motion for certain elliptically shaped tables. The motion of the ball can be described by a set of discrete numbers $(s_i,\ \phi_i)$ called a *map*.

Another example, which the reader with access to a laboratory can see for oneself, is the ball in a two-well potential shown in Fig. 11.1(b). Here the ball has two equilibrium states when the table or base does not vibrate. However, when the table vibrates with periodic motion, the jumps of the ball from one well to the other may be periodic under certain conditions, i.e., LRLR $\cdots$ or LLRLLR $\cdots$, and so on; and may be even chaotic for other conditions, i.e., it shows no pattern in the sequence of symbols L and R. A periodic input of one frequency leads to randomlike output with a broad spectrum of frequencies is one of characteristics of chaotic vibrations.

We have learned in detail one-dimensional discrete map – the logistic map in Chapter 4. In this chapter we first consider a 2-D, invertible maps – Hénon map, and then study some examples of chaos in differential equations.

11.2 Two-Dimensional Invertible Maps

A general 2-D map can be written as

$$\begin{aligned} x_{n+1} &= f_1(x_n\ y_n) \\ y_{n+1} &= f_2(x_n\ y_n)\,. \end{aligned} \tag{11.1}$$

Actually, 2-D invertible maps can be simply constructed from 1-D non-invertible maps $F(x)$ as

$$\begin{aligned} x_{n+1} &= F(x_n) + y_n \\ y_{n+1} &= \beta\, x_n\,. \end{aligned} \tag{11.2}$$

For $\beta = 0$, (11.2) gives back the 1-D non-invertible map. But as long as $\beta \neq 0$ (no matter how small) we can always invert it by

$$\begin{aligned} x_n &= \frac{y_{n+1}}{\beta} \\ y_n &= x_{n+1} - F\left(\frac{y_{n+1}}{\beta}\right). \end{aligned} \tag{11.3}$$

Since the determinant of the Jacobian of (11.2) is

$$\det \begin{bmatrix} \dfrac{\partial x_{n+1}}{\partial x_n} & \dfrac{\partial x_{n+1}}{\partial y_n} \\ \dfrac{\partial y_{n+1}}{\partial x_n} & \dfrac{\partial y_{n+1}}{\partial y_n} \end{bmatrix} = -\beta\,. \tag{11.4}$$

Hence, if $|\beta| < 1$, the phase space area contacts at a rate $\propto |\beta|$ at each iteration.

Thus, if the sequence $\{(x_n\,,\,y_n)\}$ remains in a bounded region of the place (attractor) it must asymptotically approach a set of area zero. For example, if the sequence becomes attracted to a N-point periodic cycle, then the attractor is the set of N points plotted in the $(x,\,y)$ plane.

Another property is that the attractor is a curve, which still has zero area. However, there can be attractors which have non-integer dimension. Such attractors are called *strange attractor* and the dimension of the strange attractor is called *fractal dimension*. More detail about the fractal dimension will be discussed in Chapter 12. To illustrate the properties of 2-D maps, we consider the **Hénon map**.

Hénon had a clever idea in developing a method to overcome the numerical difficulties encountered in trying to resolve the tightly packed sheets of the Lorenz attractor. The difficulties stem from the rapid volume contraction in the Lorenz system. Hénon sought a mapping that captured its essential features, and he chose to study mapping rather than differential equations because maps are faster to simulate and their solutions can be followed more accurately.

The Hénon map is given by

$$\begin{aligned} x_{n+1} &= 1 - \alpha\, x_n^2 + y_n \\ y_{n+1} &= \beta\, x_n\,, \qquad (\beta \neq 0) \end{aligned} \tag{11.5}$$

where α and β are adjustable parameters. Hénon showed that this is the most general quadratic map with constant Jacobian. There are two fixed points:

$$\begin{aligned} x^{(1)} &= \frac{1}{2\,\alpha}\left[(\beta-1) \pm \sqrt{(\beta-1)^2 + 4\,\alpha}\right] \\ y^{(1)} &= \beta\, x^{(1)} \end{aligned} \tag{11.6}$$

provided $\alpha > \alpha_0 = -\dfrac{(\beta-1)^2}{4}$. Linear stability analysis shows that $x^{(1)}$ is unstable if

$$\alpha > \alpha_1 = \frac{3}{4}(\beta-1)^2. \tag{11.7}$$

When α is increases past α_1 a series of bifurcations is observed and the attractor is still simple, consisting of a periodic set of $p = 2^n$ points with $p \to \infty$ as $\alpha \to \alpha_2$. Beyond this, for $\alpha_2 < \alpha < \alpha_3$, the numerical evidence strongly suggests the presence of a strange attractor.

We first summarize the elementary properties of the Hénon map and then give some numerical simulation results.

1) *The Hénon map is invertible.*

2) *The Hénon map is dissipative.* It contracts areas, and does so at the same everywhere in phase space.

3) *For certain parameter values, the Hénon map has a trapping region.* In other words, there is a region R that gets mapped inside itself (Fig. 11.2).

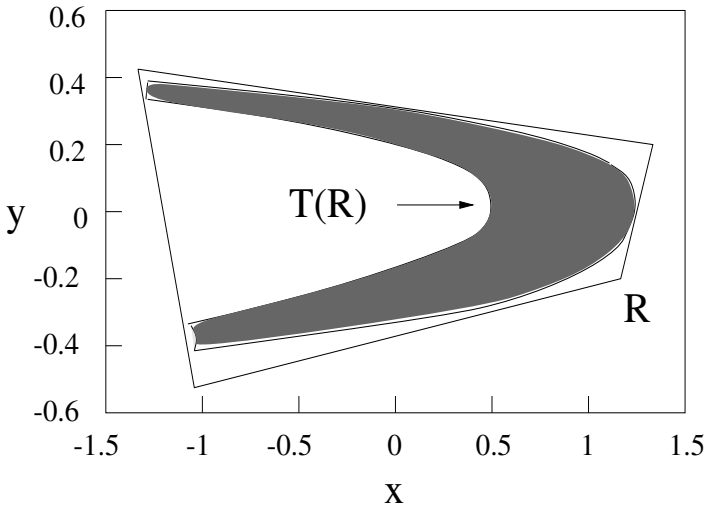


Fig. 11.2 Trapping region of the Hénon map.

4) *Some trajectories of the Hénon map escape to infinity.* This is because far from the origin, the quadratic term in (11.5) dominates and repels orbits to infinity. Similar

behavior occurs in the logistic map – recall that orbits starting outside the unit interval eventually become unbounded.

It is easy to show that the Hénon map is invertible if $\beta \neq 0$; and the Hénon map contracts areas if $|\beta| < 1$ by considering the determinant of the Jacobian.

Next, we want to find suitable values of parameters α and β to exhibit the strange attractor. First note that β should not be too close to zero, or else the area contraction will be excessive and the fine structure of the attractor will be invisible. A good choice is $\beta = 0.3$. To fine a good value of α, we need some exploring. If α is too small or too large, all trajectories escape to infinity; there is no attractor in these cases. For intermediate values of α, the trajectories either escape to infinity or approach an attractor, depending upon the initial conditions. As α increases through this range, the attractor changes from a stable fixed point to a stable 2-cycle. The system then undergoes a period-doubling route to chaos, followed by chaos intermingled with periodic windows. Hénon picked $\alpha = 1.4$, well into the chaotic region. The attractor is bent like a boomerang and is made of many parallel curves, see Fig. 11.3(a). Fig 11.3(b) is an enlargement of the small square of Fig. 11.3(a). The characteristic fine structure of the attractor begins to emerge. These seem to be six parallel curves: a lone curve near the middle of the frame, then two closely spaced curves above it, and then three more. If we zoom in on those three curves (Fig. 11.3(c)), it becomes clear that they are actually six curves, grouped one, two, three, exactly as before! And those curves are themselves made of thinner curves in the same pattern, and so on. The self-similarity continues to arbitrarily small scales.

11.3. Three-Dimensional Systems of ODEs: Lorenz Equations

Now we turn to differential equations. For differential equations to have chaos, its dimension of phase space must be at least three. In this section, we mainly discuss the Lorenz equations.

11.3.1 Lorenz equations. Ed Lorenz (1963) derived a three-dimensional chaotic system from a drastically simplified model of convection rolls in the atmosphere. This well-

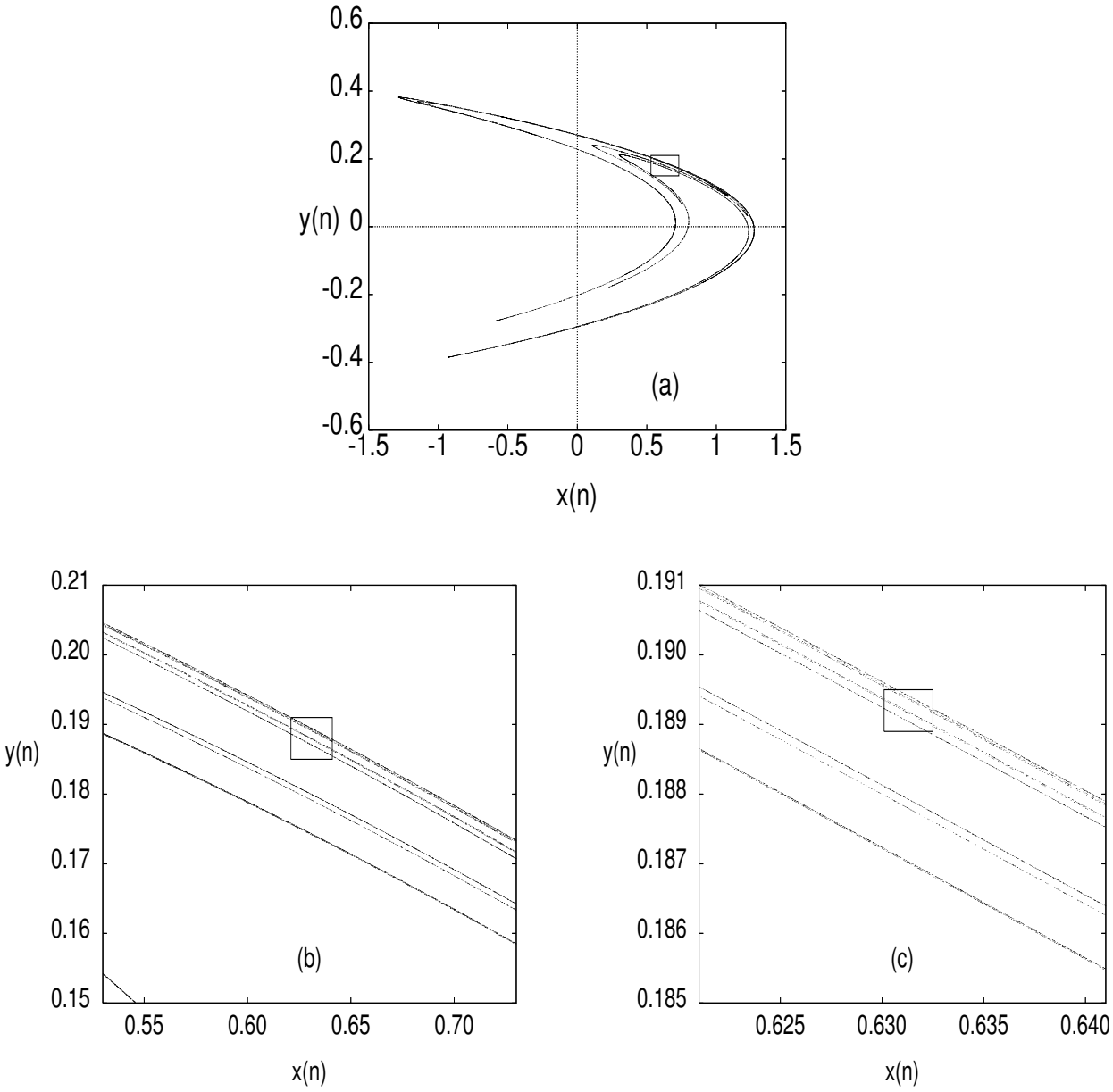


Fig. 11.3 Strange attractor of Hénon map when $\alpha = 1.4$ and $\beta = 0.3$.

known equation is now called Lorenz equations, given by

$$\begin{aligned} \dot{x} &= \sigma\,(y - x) \\ \dot{y} &= \rho\, x - y - x\, z \\ \dot{z} &= x\, y - \beta\, z \end{aligned} \tag{11.8}$$

where σ, ρ, $\beta > 0$ are parameters. The same equations also arise in models of lasers and

dynamos. Equation (11.8) can be written in the vector form $\dfrac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x})$.

Lorenz discovered that this simple-looking deterministic system could have extremely erratic dynamics: over a wide range of parameters, the solutions oscillate irregularly, never exactly repeating but always remaining in a bounded region that they settled onto a complicated set, now called a *strange attractor*. Unlike stable fixed points and limit cycles, the strange attractor is not a point or a curve or even a surface – it's a fractal, with a fractional dimension between 2 and 3.

11.3.2 Simple properties of the Lorenz equations. First note that the system (11.8) has only two nonlinearities, the quadratic terms $x\,y$ and $x\,z$. It is also noted that an important *symmetry* in the Lorenz system: if we replace $(x,\, y) \to (-\,x,\, -\,y)$ in (11.8), the equations stay the same. Hence, if $(x(t),\, y(t),\, z(t))$ is a solution, so is $(-\,x(t),\, -\,y(t),\, z(t))$.

Next, we prove that the Lorenz system is *dissipative*: volumes in phase space contract under the flow. To see this, we must first ask: how do volume evolve? We note that for $\sigma\,,\ \beta > 0\,,$ the trace of $D\boldsymbol{f}$ is always negative; that is, $\bigtriangledown \bullet \boldsymbol{f}$ is negative or $\bigtriangledown \bullet \dfrac{d\boldsymbol{x}}{dt}$ is negative. What does this have to do with volume? Well, suppose water fills a tub from some point source at rate $\varepsilon\,.$ If we imagine a surface Σ around the source, the rate at which the tub is filled is just $\int_\Sigma \rho\,\boldsymbol{v} \bullet d\sigma\,,$ where ρ is the density of water and $\boldsymbol{v}$ is the velocity of water out of Σ. Then $\varepsilon = \int_\Sigma \rho\,\boldsymbol{v} \bullet d\sigma\,,$ and so $\dfrac{\varepsilon}{\rho} = \displaystyle\int_\Sigma \boldsymbol{v} \bullet d\sigma.$ The integrand is the rate at which volume is filled by water. Suppose this source is not a point source, but a uniform source function everywhere in $\Sigma\,,$ then

$$\int_V \frac{\varepsilon}{\rho}\, dV = \int \boldsymbol{v} \bullet d\sigma \quad \Longrightarrow \quad \frac{\varepsilon}{\rho} = \bigtriangledown \bullet \boldsymbol{v} \tag{11.9}$$

due to Gauss's Law, where $\bigtriangledown \bullet \boldsymbol{v}$ is the local rate of volume filled by water. Thus, $\boldsymbol{v}$ is the volume "flux density". So, $\bigtriangledown \bullet \dfrac{d\boldsymbol{x}}{dt}$ represents the rate of growth of a volume due to the dynamical flow.

To put the above argument more rigorously, let's consider the question in general, for any three-dimensional system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$. Pick an arbitrary closed surface $S(t)$ of volume $V(t)$ in phase space. Think of the points on S as initial conditions for trajectories, and let them evolve for an infinitesimal time dt. Then S evolves into a new surface $S(t+dt)$; what

is its volume $V(t+dt)$? Fig. 11.4(a) shows a side view of the volume.

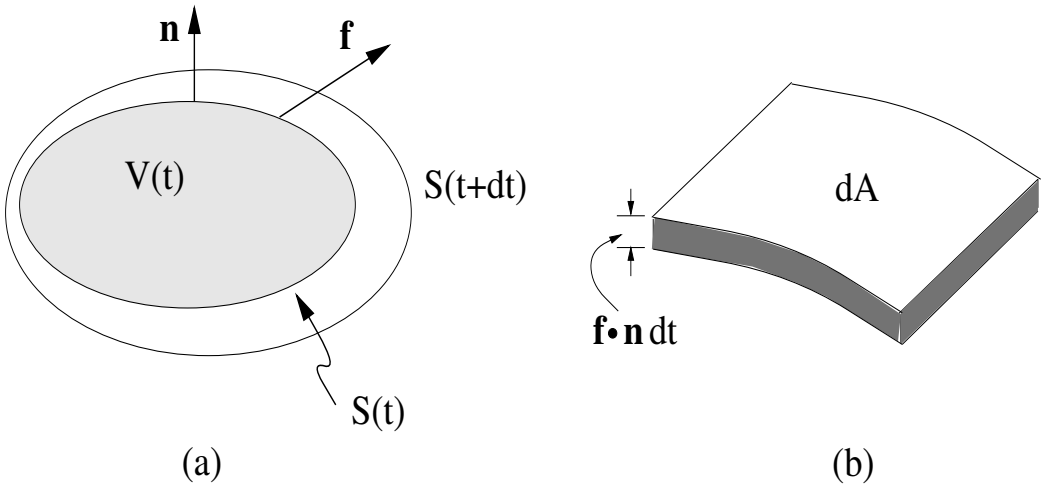


Fig. 11.4 Volume contraction.

Let $\boldsymbol{n}$ denote the outward normal on S. Since $\boldsymbol{f}$ is the instantaneous velocity of the points. $\boldsymbol{f} \bullet \boldsymbol{n}$ is the outward norm component of velocity. Therefore, in time dt a patch of area dA sweeps out a volume $(\boldsymbol{f} \bullet \boldsymbol{n}\, dt)\, dA$, as shown in Fig. 11.4(b). Hence, $V(t+dt) = V(t)+$ (volume swept out by tiny patches of surface, integrated over all patches), so we obtain

$$V(t+dt) = V(t) + \int_S (\boldsymbol{f} \bullet \boldsymbol{n}\, dt)\, dA\,.$$

Thus,

$$\dot{V} = \frac{V(t+dt) - V(t)}{dt} = \int_S \boldsymbol{f} \bullet \boldsymbol{n}\, dA\,.$$

Finally, we rewrite the above integral by the divergence theorem, and get

$$\dot{V} = \int_V \bigtriangledown \bullet \boldsymbol{f}\, dV\,. \tag{11.10}$$

For the Lorenz system,

$$\bigtriangledown \bullet \boldsymbol{f} = \frac{\partial}{\partial x}\,\sigma\,(y-x) + \frac{\partial}{\partial y}\,(\rho\,x - y - x\,z) + \frac{\partial}{\partial x}\,(x\,y - \beta\,z) = -\,\sigma - 1 - \beta < 0\,. \tag{11.11}$$

Since the divergence is constant, (11.10) reduces to $\dot{V} = -\,(\sigma - 1 - \beta)\,V$, which has solution $V(t) = V(0)\,e^{-\,(\sigma+1+\beta)\,t}$. Thus *volumes in phase space shrink exponentially fast.*

Volume contraction imposes strong constraints on the possible solutions of the Lorenz equations. It can be shown that there are no quasiperiodic solutions of the Lorenz equations. Because if there were a quasiperiodic solution, it would have to lie on the surface of a torus, and this torus would be *invariant* under the flow. Hence, the volume inside the torus would be constant in time. But this contradicts the fact that all volumes shrink exponentially fast. Also, we show that it is impossible for the Lorenz system to have either repelling fixed points or repelling closed orbits. Repellers are incompatible with volume contraction because they are *sources* of volume, in the following sense. Suppose we encase a repeller with a closed surface of initial conditions nearby in phase space. A short time later, the surface will have expanded as the corresponding trajectories are driven away. Thus, the volume inside the surface would increase. This contradicts the fact that all volumes contract.

Therefore, we conclude that all fixed points of the Lorenz system must be sinks or saddles, and closed orbits (if they exist) must be stable or saddle-like. For the case of fixed points, we now verify these general conclusions explicitly.

11.3.3 Fixed points and their stability. The Lorenz system (11.8) has two types of fixed points. The origin

$$(x^*,\, y^*,\, z^*) = (0,\, 0,\, 0) \tag{11.12}$$

is a fixed point for *all* values of the parameters. For $\rho > 1$, there is also a symmetric pair of fixed points

$$\begin{aligned} x^* = y^* &= \pm\sqrt{\beta\,(\rho-1)} \\ z^* &= \rho - 1\,. \end{aligned} \tag{11.13}$$

Lorenz called them C^+ and C^-. Then represent left- or right-turning convection rolls. As $\rho \to 1^+$, C^+ and C^- coalesce with the origin in a *pitchfork* bifurcation.

The linearization of the Lorenz system at the origin is obtained by omitting the $x\,y$ and $x\,z$ nonlinear terms in (11.8). The equation for z is decoupled and shows that $z(t) \to 0$ exponentially fast. The other two directions are governed by the system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} \begin{bmatrix} -\sigma & \sigma \\ \rho & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{11.14}$$

with trace $\tau = -\sigma - 1 < 0$ and determinant $\Delta = \sigma\,(1-\rho)$. If $\rho > 1$, the origin is a saddle point because $\Delta < 0$. Note that this is a *new type of saddle*, since the full system is three dimension. Including the decaying z-direction, the saddle has one outgoing and two incoming directions. If $\rho < 1$, all directions are incoming and the origin is a sink. Specifically, since $\tau^2 - 4\,\Delta = (\sigma+1)^2 - 4\,\sigma\,(1-\rho) = (\sigma-1)^2 + 4\,\sigma\,\rho > 0$, the origin is a stable node for $\rho < 1$.

Actually, we can prove that the origin is *globally stable* when $\rho < 1$. Consider the Liapunov function

$$V(x,\,y,\,z) = \frac{1}{\sigma}\,x^2 + y^2 + z^2 \tag{11.15}$$

which is positive definite for $(x,\,y,\,z) \neq (0,\,0,\,0)$; and

$$\begin{aligned}
\frac{1}{2}\,\dot{V} &= \frac{1}{\sigma}\,x\,\dot{x} + y\,\dot{y} + z\,\dot{z} \\
&= (y\,x - x^2) + (\rho\,y\,x - y^2 - x\,z\,y) + \left(z\,x\,y - \beta\,z^2\right) \\
&= (\sigma+1)\,x\,y - x^2 - y^2 - \beta\,z^2 \\
&= -\left[x - \frac{\sigma+1}{2}\,y\right]^2 - \left[1 - \left(\frac{\sigma+1}{2}\right)^2\right] y^2 - \beta\,z^2
\end{aligned} \tag{11.16}$$

which shows that $\dot{V}$ is strictly negative if $\rho < 1$ and $(x,\,y,\,z) \neq (0,\,0,\,0)$. Therefore, the origin is globally stable for $\rho < 1$ and *every* trajectory approaches the origin as $t \to \infty$.

Now suppose $\rho > 1$, so that C^+ and C^- exist. We can use the following notations:

$$C^+ :\ (x_+\,,\ x_+\,,\ \frac{1}{\beta}\,x_+^2) \qquad \text{and} \qquad C^- :\ (x_-\,,\ x_-\,,\ \frac{1}{\beta}\,x_-^2) \tag{11.17}$$

where $x_\pm = \pm\sqrt{\beta\,(\rho-1)}$. The linearization yields the characteristic polynomial

$$\begin{aligned}
& -(\lambda+\beta)\,\left[\lambda^2\,(1+\sigma)\,\lambda + \sigma - \sigma\,(\rho - z_\pm)\right] + x_\pm\,\left[\,-(\sigma+\lambda)\,x_\pm - \sigma\,y_\pm\,\right] = 0 \\
\Longrightarrow \\
& (\lambda+\beta)\,\left[\lambda^2 + (\sigma-1)\,\lambda\right] + x_\pm^2\,(2\,\sigma+\lambda) = 0 \\
\Longrightarrow \\
& (\lambda+\beta)\,\left[\lambda^2 + (\sigma_1)\,\lambda\right] + \beta\,(\rho-1)\,(2\,\sigma+\lambda) = 0 \\
\Longrightarrow \\
& \lambda^3 + (\sigma+\beta+1)\,\lambda^2 + \beta\,(\sigma+\rho)\,\lambda + 2\,\sigma\,\beta\,(\rho-1) = 0
\end{aligned} \tag{11.18}$$

By using the Hurwitz criterion, stability of the fixed points C^+ and C^- requires that all the three coefficients in the above polynomial must be positive, which yields $\rho > 1$, and in addition,

$$\begin{aligned} & (\sigma + \beta + 1)\,\beta\,(\sigma + \rho) - 2\,\sigma\,\beta\,(\rho - 1) > 0 \\ \Longrightarrow & \\ & \beta\,[\,(\sigma + \beta + 3)\,\sigma - (\sigma - \beta - 1)\,\rho\,] > 0 \\ \Longrightarrow & \\ & (\sigma + \beta + 3)\,\sigma - (\sigma - \beta - 1)\,\rho > 0\,. \end{aligned} \tag{11.19}$$

Therefore, when $(\sigma + \beta + 3)\,\sigma = (\sigma - \beta - 1)\,\rho$, i.e., at the critical point

$$\rho_H = \frac{\sigma + \beta + 3}{\sigma - \beta - 1}\,\sigma \tag{11.20}$$

(assuming also that $\sigma - \beta - 1 > 0$), a Hopf bifurcation occurs. Here the subscript H indicates the Hopf bifurcation.

What happens immediately after the bifurcation, for ρ slightly greater than ρ_H? One might guess that C^+ and C^- would each be surrounded by a small stable limit cycle. That would occur if the Hopf bifurcation were supercritical. But actually it's *subcritical* – the limit cycles are *unstable* and exist only for $\rho < \rho_H$.

Here is the intuitive picture. For $\rho < \rho_H$ the phase portrait near C^+ is shown schematically in Fig. 11.5. The fixed point is stable. It is encircled by a *saddle cycle*, a new type of unstable limit cycle that is possible only in phase spaces of three or more dimensions. The cycle has a two-dimensional unstable manifold (the sheet in Fig. 11.5), and a two-dimensional stable manifold (not shown). As $\rho \to \rho_H$ from below, the cycle shrinks down around the fixed point. At the Hopf bifurcation, the fixed point absorbs the saddle cycle and changes into a saddle point. For $\rho > \rho_H$ there are no attractors in the neighborhood.

So for $\rho > \rho_H$ trajectories must fly away to a distant attractor. But what can it be? A partial bifurcation diagram for the system, based on the results so far, shows no hint of any stable objects for $\rho > \rho_H$ (Fig. 11.6).

Could it be that all trajectories are repelled out to infinity? No; we can prove that all trajectories eventually enter and remain in a certain large ellipsoid. Could there be some

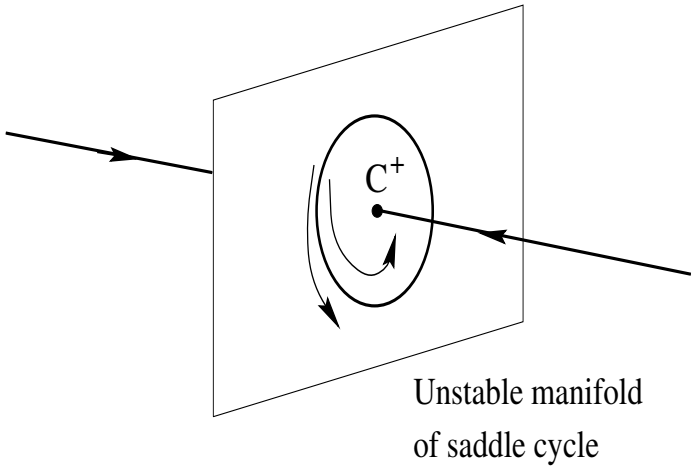


Fig. 11.5 Saddle cycle.

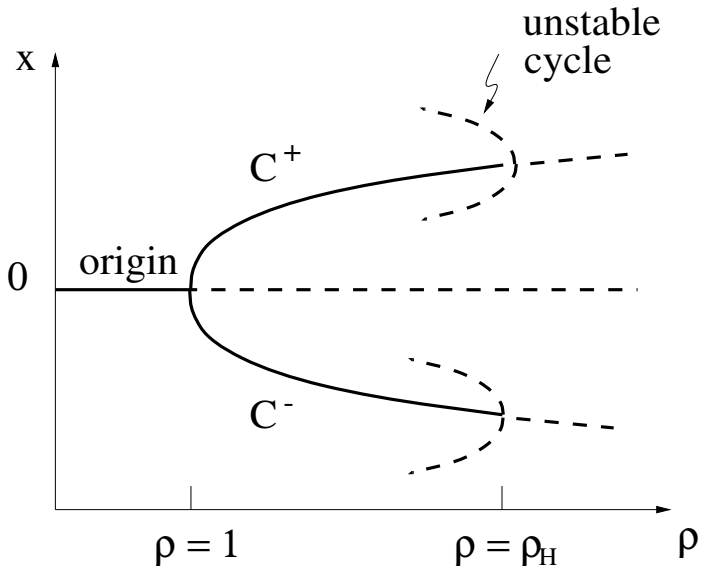


Fig. 11.6 Bifurcation diagram.

stable limit cycles we're unaware of? Possibly, but Lorenz gave a persuasive argument that for ρ slightly greater than ρ_H, any limit cycles would have to be unstable.

So the trajectories must have a bizarre kind of long-term behavior. Like balls in a pinball machine, they are repelled from one unstable object after another. At the same time, they are confined to a bounded set of zero volume, yet they manage to move on this set forever without intersecting themselves or others.

11.3.4 Chaos on a strange attractor. Lorenz used numerical integration to see what the trajectories would do in the long run. He studied the particular case: $\sigma = 10$, $\beta = \dfrac{8}{3}$, $\rho = 28$. This value of ρ is just past the Hopf bifurcation value $\rho_H = \dfrac{\sigma\,(\sigma + \beta + 3)}{\sigma - \beta - 1} \approx 24.74$, so he knew that something strange had to occur. The Fig. 11.7 shows the time history of $y(t)$ starting from the point $(0,\ 1,\ 0)$. After an initial transient, the solution settles into an irregular oscillation that persists as $t \to \infty$, but never repeats exactly. The motion is *aperiodic*. Lorenz discovered that a wonderful structure emerges if the solution is visualized as a trajectory phase space. For instance, when $x(t)$ is plotted against $z(t)$, a butterfly pattern appears (Fig. 11.8).

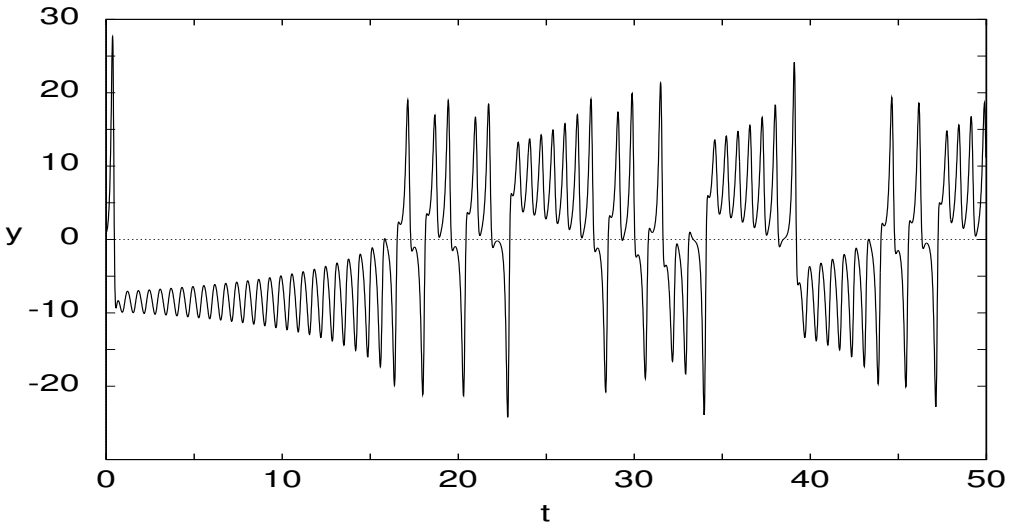


Fig. 11.7 Time history of $y(t)$.

The trajectory appears to cross itself repeatedly, gut that's just an artifact of projecting the three-dimensional trajectories onto a two-dimensional plane. In three dimensions no self-intersections occur. Two other different projections of the strange attractor are shown in Fig. 11.9.

Let's try to understand Fig. 11.8 in detail. The trajectory starts near the origin, then swing to the right, and then dives into the center of a spiral on the left. After a very slow spiral outward, the trajectory shoots back over to the right side, spirals around a few times, shoots over to the left, spirals around, and so on indefinitely. The number of circuits made

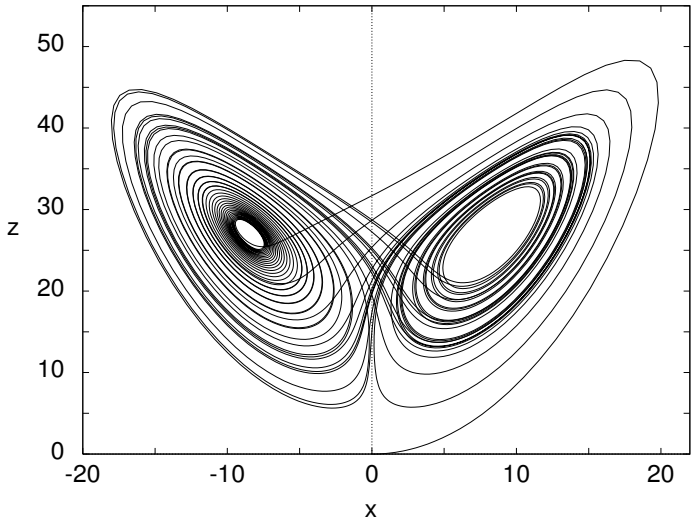


Fig. 11.8 Butterfly of the Lorenz system.

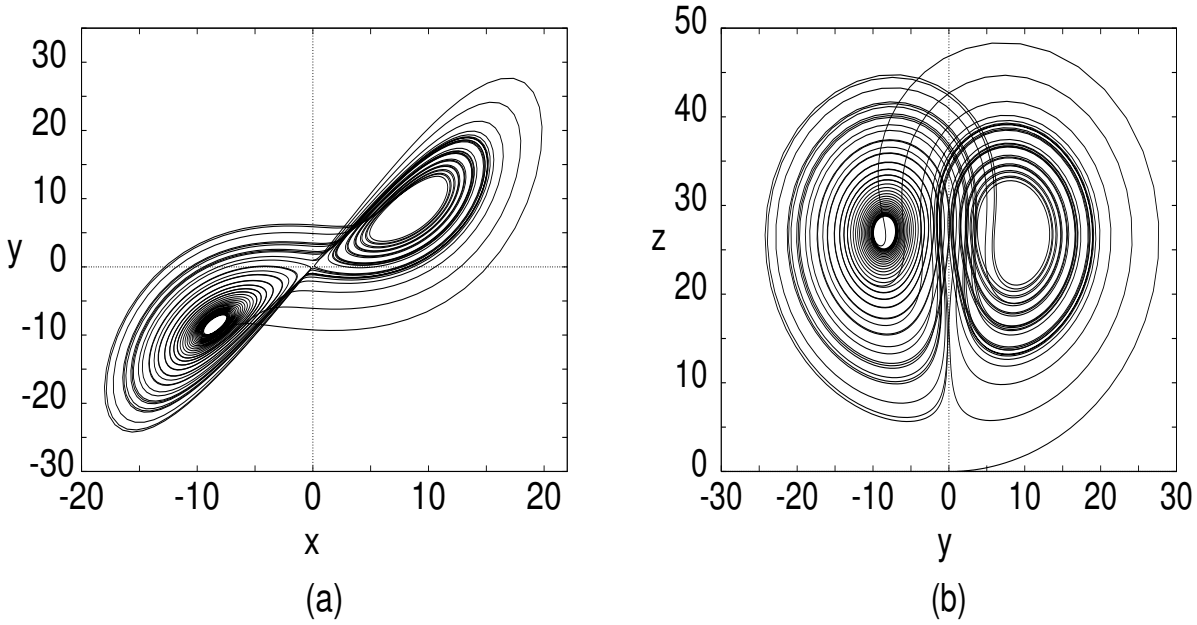


Fig. 11.9 Another view of the Lorenz attractor.

on either side varies unpredictably from one cycle to the next. In fact, the sequence of the number of circuits has many of the characteristics of *random sequence*.

When the trajectory is viewed in all three dimensions, rather than in a two-dimensional projection, it appears to settle onto an exquisitely thin set that looks like a pair of butterfly

wing. Fig. 11.10 shows a schematic of the **strange attractor** (a term coined by Rulle and Takens (1971)). This limiting set is the attracting set of zero volume whose existence was deduced before.

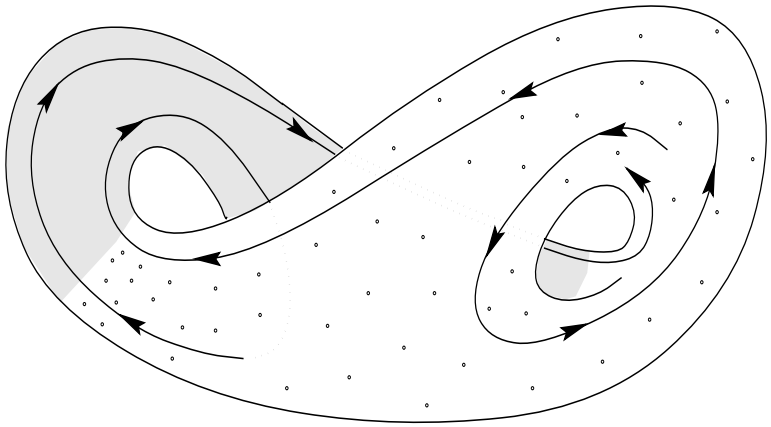


Fig. 11.10 Strange attractor.

What is the geometrical structure of the strange attractor? Fig. 11.10 suggests that it is pair of surfaces that merge into one in the lower portion of Fig. 11.10. But how can this be, when the uniqueness theorem tells us that trajectories can't cross or merge? Lorenz (1963) gives a lovely explanation – the two surfaces only *appear to merge*. The illusion is caused by the strong volume contraction of the flow, and insufficient numerical resolution. But when watch where that idea lead him:

> It would seem, then, that the two surfaces merely appear to merge, and remain distinct surfaces. Following these surfaces along a path parallel to a trajectory, and circling C^+ and C^-, we see that each surface is really a pair of surfaces, so that, where they appear to merge, there are really four surfaces. Continuing this process for another circuit, we see that there are really eight surfaces, etc., and we finally conclude that there is an infinite complex of surfaces, each extremely close to one or other of two merging surfaces.

Today this "infinite complex of surfaces" would be called a fractal. It is a set of points with zero volume but infinite surfaces area. In fact, numerical experiments suggest that it has a

dimension of about 2.05!

11.3.5 Definition of chaos. No definition of the term *chaos* is universally accepted yet, but almost everyone would agree on the three ingredients used in the following working definition: **Chaos** is *aperiodic long-term behavior* in a deterministic system that exhibits *sensitive dependence on initial conditions.*

1) *Aperiodic long-term behavior* means that there are trajectories which do not settle down to fixed points, periodic orbits, or quasiperiodic orbits as $t \to \infty$. For practical reasons, we should require that such trajectories are not too rare. For instance, we could insist that there be an open set of initial conditions leading to aperiodic trajectories, or perhaps that such trajectories should occur with nonzero probability, given a random initial condition.

2) *Deterministic* means that the system has no random or noisy inputs or parameters. The irregular behavior arises from the system's nonlinearity, rather than from noisy driving forces.

3) *Sensitive dependence on initial condition* means that nearby trajectories separate exponentially fast, i.e., the system has a positive Liapunuv exponent.

The term *attractor* is also difficult to define in a rigorous way. Loosely speaking, an attractor is a set to which all neighboring trajectories converge. Stable fixed points and stable limit cycles are examples. Finally, we define a **strange attractor** to be an attractor that exhibits sensitive dependence on initial conditions. Strange attractors were originally called strange because they are often fractal sets. Nowadays this geometric property is regarded as less important than the dynamical property of sensitive dependence on initial conditions. The term *chaotic attractor* and *fractal attractor* are used when one wishes to emphasize one or the other of those aspects.

11.3.6 Transient chaos. Consider the Lorenz equations numerically for the parameter values $\sigma = 10$, $\beta = \dfrac{8}{3}$ and $\rho = 21$. Note that under these parameter values, the fixed points C^+ and C^- are still asymptotically stable. After experimenting with a few different initial

conditions, it is easy to find solutions like that shown in Fig. 11.11 where the initial conditions are chosen as $(x,\, y,\, z) = (10,\, 0,\, 5)$. At first the trajectory seems to be tracing out at strange attractor, but eventually it stays on the left and spiral down toward the stable fixed point C^-. The time series of y vs. t shows the same result: an initially erratic solution ultimately damps down to equilibrium (Fig. 11.12).

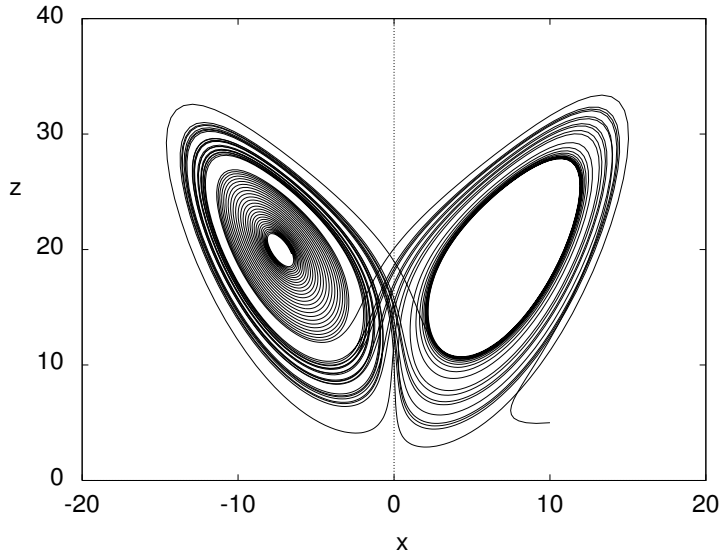


Fig. 11.11 A trajectory of the Lorenz system when $\sigma = 10,\ \beta = \dfrac{8}{3},\ \rho = 21.$

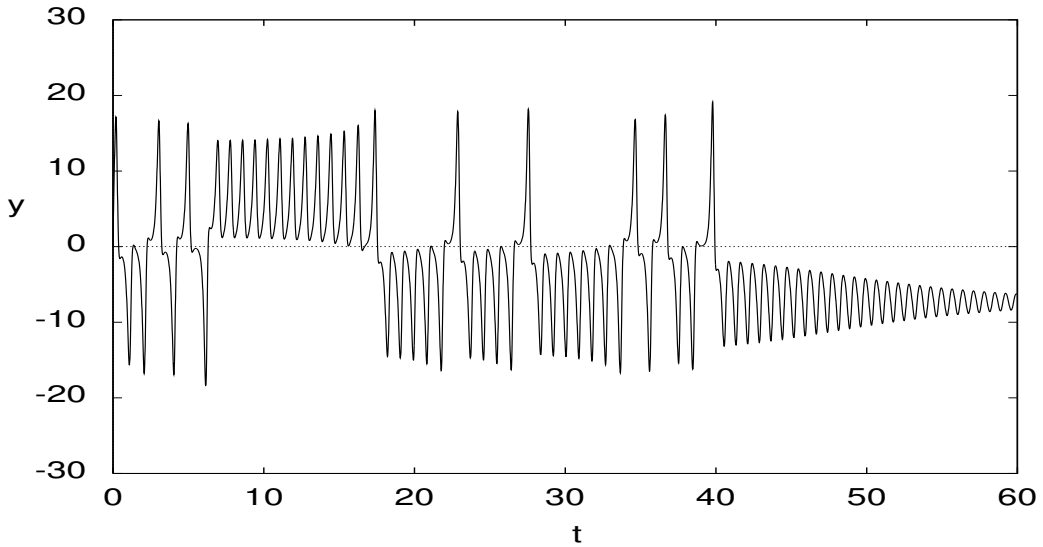


Fig. 11.12 Time history of $y(t)$.

Other names used for transient chaos are *metastable chaos* (Kaplan and Yorke, 1979) or *pre-turbulence* (Yorke and Yorke, 1979; Sparrow, 1982).

By our definition, the above dynamics are not "chaotic," because the long-term behavior is not aperiodic. On the other hand, the dynamics do exhibit sensitive dependence on initial conditions – if we have chosen a slightly different initial condition, the trajectory could easily have ended up at C^+ instead of C^-. Thus the system's behavior is unpredictable, at least for certain conditions.

Transient chaos shows that a deterministic system can be unpredictable, even its final states are very simple. In particular, you don't need strange attractors to generate effectively random behavior. Of course, this is familiar from everyday experience – many games of "chaos" used in gambling are essentially demonstrations of transient chaos. For instance, think about rolling dice. A crazily-rolling die always stops in one of six stable equilibrium positions. The problem with predicting the outcome is that the final position depends sensitively on the initial orientation and velocity (assuming the initial velocity is large enough).

11.3.7 Using chaos to send secret message. One of the most exciting recent developments in nonlinear dynamics is the realization that chaos can be *useful*. Normally one thinks of chaos as a fascinating curiosity at best, and a nuisance at worst, something to be avoided or engineered away. But since about 1990, people have found ways to exploit chaos to do some marvelous and practical things. For an introduction to this new subject, see Vohra et al. (1992).

One application involves "private communications." Suppose we want to send a secret message to a friend or business partner. Naturally we should use a code, so that even if an enemy is eavesdropping, he will have trouble making sense of the message. This is an old problem – people have been making (and breaking) codes for as long as they have been secrets worth keeping.

Kevin Cuomo and Alan Oppenheim (1992, 1993) have implemented a new approach to this problem, building on Pecora and Carroll's (1990) discovery of **synchronized chaos**. Here's the strategy: When we transmit the message to our friend, as also "mask" it with much louder chaos. An outside listener only hears the chaos, which sounds like meaningless

noise. But now suppose that our friend has a magic receiver that perfectly reproduces the chaos – then he can subtract off the chaotic mask and listen to the message!

The fundamental idea behind this method is that the transmitter and receiver are made of an identical chaotic circuit system, and these two circuits can be synchronized after a transient period no matter what the initial conditions are. Next we prove this synchronization. Write down the governing equations for both the transmitter and the receiver.

$$\begin{aligned} \dot{x} &= \sigma\,(y - x) \\ \dot{y} &= \rho\, x - y - x\, z \\ \dot{z} &= x\, y - \beta\, z \end{aligned} \tag{11.21}$$

for the transmitter and

$$\begin{aligned} \dot{x}_r &= \sigma\,(y_r - x_r) \\ \dot{y}_r &= \rho\, x(t) - y_r - x(t)\, z_r \\ \dot{z}_r &= x(t)\, y_r - \beta\, z_r \end{aligned} \tag{11.22}$$

for the receiver, where we have written $x(t)$ to emphasize that the receiver is driven by the chaotic signal $x(t)$ coming from the transmitter. This process can be schematically explained in Fir. 7.36.

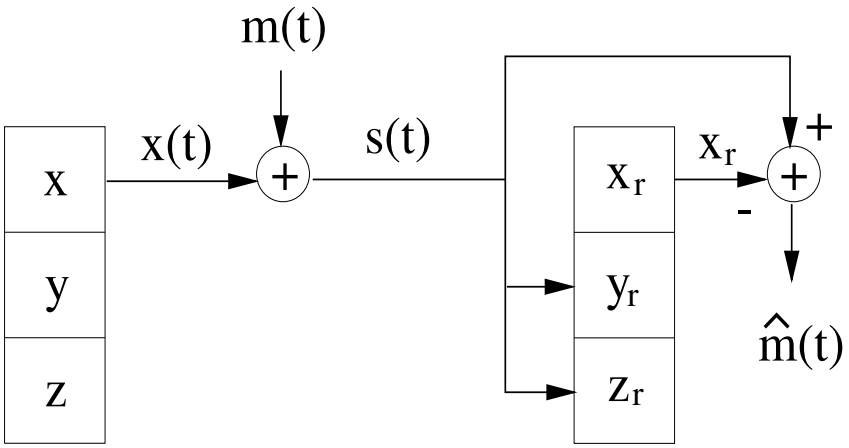


Fig. 11.13 Chaotic signal masking system.

The astonishing result is that *the receiver asymptotically approaches perfect synchrony with the transmitter, starting from any initial condition*! To be precise, let

$\boldsymbol{d} = (x,\, y,\, z) =$ state of the transmitter or "driver"

$\boldsymbol{r} = (x_r,\, y_r,\, z_r) =$ state of the receiver

$\boldsymbol{e} = \boldsymbol{d} - \boldsymbol{r}$ error signal.

Suppose $m(t) = 0$ (no message). We claim that $\boldsymbol{e} \to \boldsymbol{0}$ as $t \to \infty$, for all initial conditions.

Why is this astonishing? Because at each instant the receiver has only *partial* information about the state of the transmitter – it is driven solely by $x(t)$, yet somehow it manages to reconstruct the other two transmitter variables $y(t)$ and $z(t)$ as well.

The equations about the error signal $\boldsymbol{e}$ are given by

$$\begin{aligned} \dot{e}_1 &= \sigma\,(e_2 - e_1) \\ \dot{e}_2 &= -\,e_2 - x(t)\,e_3 \\ \dot{e}_3 &= x(t)\,e_2 - \beta\,e_3\,. \end{aligned} \tag{11.23}$$

This is a linear system, but it has a chaotic time-dependent coefficient $x(t)$ in two terms. The idea is to construct a Liapunov function in such a way that *the chaos cancels out.* Here's how: Multiply the second equation by e_2 and the third by e_3 and add together. Then

$$\begin{aligned} e_2\,\dot{e}_2 + e_3\,\dot{e}_3 &= -\,e_2^2 - x(t)\,e_2\,e_3 + x(t)\,e_2\,e_3 - \beta\,e_3^2 \\ &= -\,e_2^2 - \beta\,e_3^2 \end{aligned} \tag{11.24}$$

and so the chaotic term disappears!

The left-hand side of equation (11.24) is $\dfrac{1}{2}\dfrac{d}{dt}\left(e_2^2 + e_3^2\right)$. This suggests the form of a Liapunov function

$$V(\boldsymbol{e},\, t) = \frac{1}{2}\left(\frac{1}{\sigma}\,e_1^2 + e_2^2 + e_3^2\right). \tag{11.25}$$

V is certainly positive definite (for $\sigma > 0$). Furthermore,

$$\begin{aligned} \dot{V} &= \frac{1}{\sigma}\,e_1\,\dot{e}_1 + e_2\,\dot{e}_2 + e_3\,\dot{e}_3 \\ &= -\,(e_1^2 - e_1\,e_2) - e_2^2 - \beta\,e_3^2 \\ &= -\,(e_1 - \frac{1}{2}\,e_2)^2 - \frac{3}{4}\,e_2^2 - \beta\,e_3^2\,. \end{aligned} \tag{11.26}$$

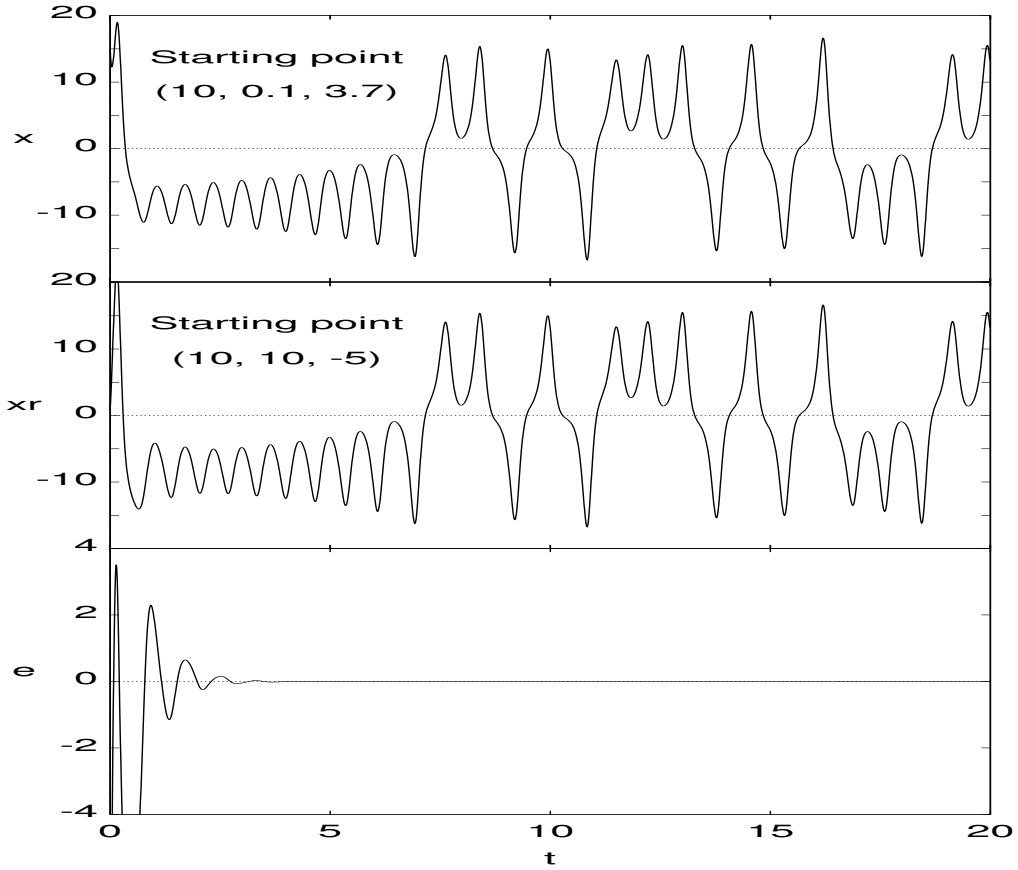


Fig. 11.14 Synchronization of the receiver to the transmitter.

Hence, $\dot{V} \leq 0$, with equality only if $\boldsymbol{e} = 0$. Therefore, V is a strict Liapunov function, and so $\boldsymbol{e}$ is globally asymptotically stable. The computer simulation results are given in Fig. 11.14, where the top, middle and bottom diagrams show the signals $x(t)$, $x_r(t)$ and $e_x(t) = x(t) - x_r(t)$, respectively.

Remark: The initial conditions used in the computer simulation (Fig. 11.14) for the transmitter and receiver are different. It is $(10.0,\ 0.1,\ 3.7)$ for the transmitter but $(10.0,\ 10.0,\ -5.0)$ for the receiver. It is amazing to see that these two chaotic systems can be asymptotically synchronized even for different initial conditions.

A stronger result is possible: one can show that $\boldsymbol{e}(t)$ decays *exponentially fast*. This is important, because rapid synchronization is necessary for the desired application.

It should be clear about what we have haven't proven. We have only shown that the

receiver will synchronize to the transmitter if the drive signal is $x(t)$. This does *not* prove that the signal-masking approach will work. For that application, the drive is a mixture $x(t) + m(t)$ where $m(t)$ is the message and $x(t) \gg m(t)$ is the mask. We have no proof that the receiver will regenerate $x(t)$ precisely. In fact, it doesn't. So it's still something of a mathematical mystery as to why the approach works as well as it does.

11.3.8 Rössler system. To end this section, we simply introduce an even simpler chaotic system than the Lorenz system – the **Rössler system**, described by

$$\begin{aligned} \dot{x} &= -y - x \\ \dot{y} &= x + \alpha\, y \\ \dot{z} &= \beta + z\,(x - \gamma) \end{aligned} \tag{11.27}$$

which has only one quadratic nonlinearity $x\,z$. Numerical integration shows that this system has a strange attractor for the parameter values: $\alpha = \beta = 0.2,\ \gamma = 5.7$ (Fig. 11.15). A schematic version of the attractor is shown in Fig. 11.16. Neighboring trajectories separate by spiraling out ("stretching"), then cross without intersecting *by going into the third dimension* ("folding") and then circulate back near their starting places ("re-injection"). We can now see why three dimensions are needed fro a flow to be chaotic.

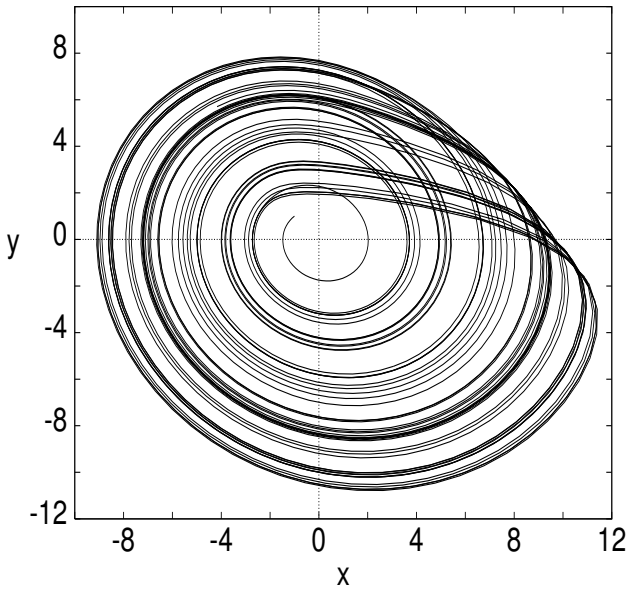


Fig. 11.15 Strange attractor of Rössler system.

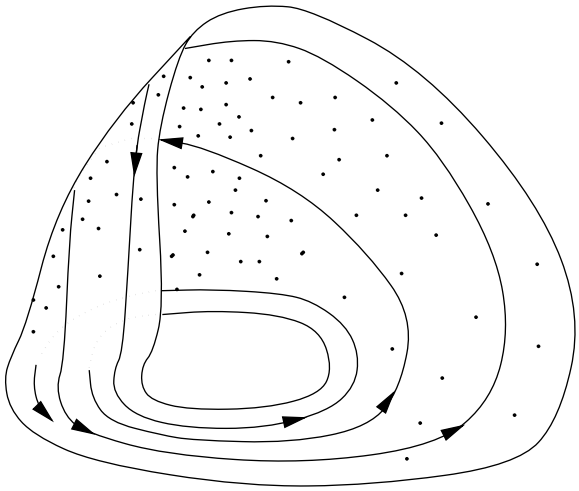


Fig. 11.16 The view of the three dimension strange attractor of Rössler system.

Chapter 12

INTRODUCTION TO FRACTALS

12.1 Introduction

In the previous chapter, we found that the solutions of Lorenz system exhibit a complicated dynamic pattern – strange attractor. As Lorenz (1963) realized, the geometry of this pattern must be very peculiar, something like an "infinite complex of surfaces." In this chapter we develop the ideas needed to describe such patterns more precisely. The tools come from fractal geometry.

Both "chaotic" and "strange attractor" have been used to describe the nonperiodic, randomlike motions of a dynamic system. Whereas "chaotic" is meant to convey a loss of information or loss of predictability, the term "strange" is meant to describe the unfamiliar geometric structure on which the motion moves in phase space. Roughly speaking, **fractals** are complex geometric shapes with fine structure at arbitrarily small scales. Usually they have some degree of self-similarity. In other words, it we magnify a tiny part of a fractal, we will see features reminiscent of the whole. Sometimes the similarity is exact; more often it is only approximate or statistical.

Fractals are of great interest because of their exquisite combination of beauty, complexity, and endless structure. They are reminiscent of natural objects like mountains, clouds, coastlines, blood vessel networks, and even broccoli, in a way that classical shapes like cones and squares can't match. They have also turned out to be useful in scientific applications ranging from computer graphics and image compression to the structural mechanics of cracks and the fluid mechanics of viscous fingering.

At the beginning of this course, we mentioned that the revolution in nonlinear dynamics has been sparked by the introduction of new geometric, analytic, and topological ideas. This in some ways parallels the earlier Newtonian revolution which introduced the calculus into dynamics. Thus, in some sense, we are entering the second phase of the Newtonian revolution

in dynamics, and new geometric concepts like fractals must be mastered if one is to use the results of the new dynamics in practical problems.

In Chapter 4, we described a quantitative measure of the chaotic aspect of these motions using Liapunov exponents. In this chapter, we will describe a quantitative measure of the strangeness of the attractor. This measure is called the *fractal dimension*. In addition to the application of fractal ideas to the description of the attractor itself, it has been discovered that other geometric objects in the study of chaos, such as the boundary between chaotic and periodic motions in initial condition or parameter space, may also have fractal properties.

Our goals in this chapter are modest. We want to become familiar with the simplest fractals and to understand the various notions of fractal dimension. Unfortunately, we will not be able to discuss the scientific applications of fractals, nor the lovely mathematical theory behind them. For the clearest introduction to the theory and applications of fractals, see Falconer (1990). The books of Mandelbrot (1982), Peitgen and Richter (1986), Barnsley (1988), Feder (1988), and Schroeder (1991) are also recommended for their many fascinating pictures and examples.

12.2 Examples of Fractals

In this section, we will discuss couple of simple examples of fractal sets and curves, namely, *Cantor sets* and *von Koch curve*. We will also introduce a quantitative measure of fractal qualities: the fractal dimension.

12.2.1 Cantor sets. How to compare two infinite sets? In the late 1800's, George Cantor invented a clever way to compare different infinite sets. Two sets X and Y are said to have the same *cardinality* (or number of elements) if there is an invertible mapping that pairs each element $x \in X$ with precisely one $y \in Y$. Such a mapping is called a **one-to-one correspondence**.

A familiar infinite set is the set of natural numbers $\boldsymbol{N} = \{1, 2, 3, 4, \cdots\}$. This set provides a basis for comparison – if another set X can be put into one-to-one correspondence with the natural numbers, then X is said to be **countable**. Otherwise X is **uncountable**. For example, even (odd) natural numbers and integers are countable.

EXAMPLE 12.1. Show that the positive rational numbers are countable.

Solution: Here's a wrong way: we start listing the numbers $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots$ in order. Unfortunately we never finish the $\frac{1}{n}$'s and so numbers like $\frac{2}{3}$ are never counted!

The right way is to make a table where the pq-th entry is $\frac{p}{q}$. Then the rationales can be counted by the weaving procedure shown in Fig. 12.1. Any given $\frac{p}{q}$ is reached after a finite number of steps, so the rationales are countable.

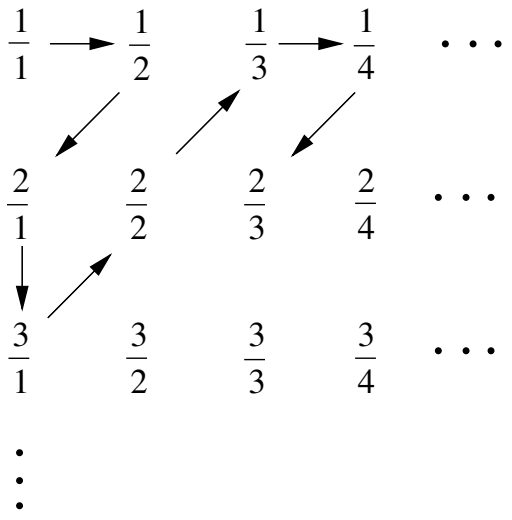


Fig. 12.1 Countable rational numbers.

EXAMPLE 12.2. Let X be the set of all real numbers between 0 and 1. Show that X is uncountable.

Solution: The proof is by contradiction. If X were countable, we could list all the real numbers between 0 and 1 as a set $\{x_1, x_2, x_3, \cdots\}$. Rewrite these numbers in decimal form:

$$\begin{aligned} x_1 &= 0.x_{11}\, x_{12}\, x_{13}\, x_{14} \cdots \\ x_2 &= 0.x_{21}\, x_{22}\, x_{23}\, x_{24} \cdots \\ x_3 &= 0.x_{31}\, x_{32}\, x_{33}\, x_{34} \cdots \\ &\vdots \end{aligned}$$

where x_{ij} denotes the jth digit of the real number x_j.

To obtain a contradiction, we'll show that there is a number α between 0 and 1 that is *not* on the list. Hence, any list is necessarily incomplete, and so X is uncountable.

We construct α as follows: its first digit is *anything other than* x_{11}, the first digit of x_1. Similarly, its second digit is anything other than the second digit of x_2. In general, the nth digit of α is $\bar{x}_{nn}$, defined as any digit other than x_{nn}. Then we claim that the number $\alpha = 0.\bar{x}_{11}\,\bar{x}_{22}\,\bar{x}_{33}\cdots$ is not on the list. Why not? It can't be equal to x_1, because it differs from x_1 in the first decimal place. Similarly, α differs from x_2 in the second decimal place, from x_3 in the third decimal place, and so on. Hence, α is not on the list, and thus X is uncountable.

This argument (devised by Cantor) is called the **diagonal argument**, because α is constructed by changing the diagonal entries x_{nn} in the matrix of digits $[x_{ij}]$.

We now turn to another of Cantor's creations, a fractal known as the Cantor set. It is simple and therefore pedagogically useful, but it is also much more than that – the Cantor set is intimately related to the geometry of strange attractors.

Fig. 12.2 shows how to construct the Cantor set.

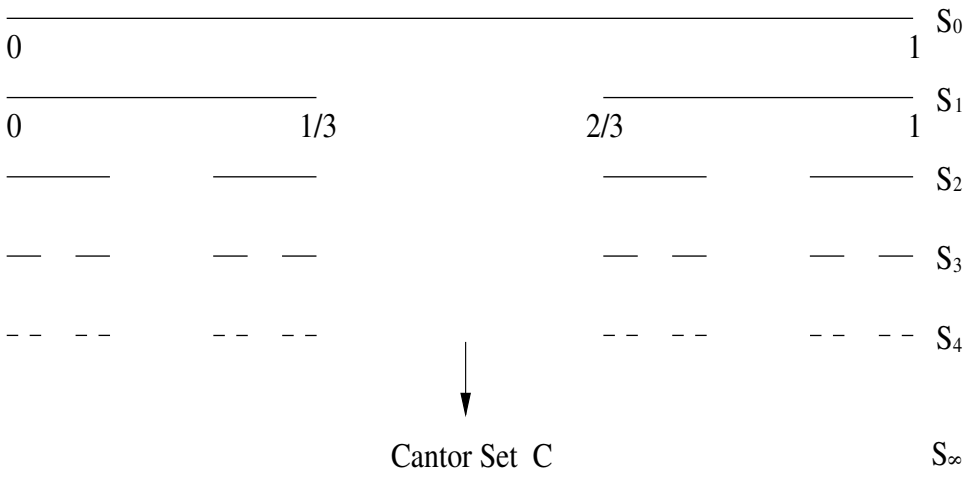


Fig. 12.2 Cantor set C.

We start with the closed interval $S_0 = [0, 1]$ and remove its open middle third, i.e., we denote the interval $(\frac{1}{3}, \frac{2}{3})$ and leave the endpoints behind. This produces the pair of closed intervals shown as S_1. Then we remove the open middle thirds of *those* two intervals to

produce S_2, and so on. The limiting set $C = S_\infty$ is the **Cantor set**. It is difficult to visualize, but Fig. 12.2 suggests that it consists of an infinite number of infinitesimal pieces, separated by gaps of various sizes.

The Cantor set C has several properties that are typical of fractal more generally:

1) *C has structure at arbitrarily small scales.* If we enlarges part of C repeatedly, we continue to see a complex pattern of points separated by gaps of various sizes. This structure is neverending. In contrast, when we look at a smooth curve or surface under repeated magnification, the picture becomes more and more featureless.

2) *C is self-similar.* It contains smaller copies of itself at all scales. For instance, if we take the left part of C (the part contained in the interval $[0,\ 1/3]$) and enlarge it by a factor of three, we get C back again. Similarly, the parts of C in each of the four intervals of S_2 are geometrically similar to C, except scaled down by a factor of nine.

   Note: The strict self-similarity of the Cantor set is found only in the simplest fractals. More general fractals are only approximately self-similar.

3) *The dimension of C is not an integer.* As we'll see later that its dimension is actually $\dfrac{\ln 2}{\ln 3} \approx 0.63$! The noninteger dimension is a natural generalization of out intuitive ideas about dimension, and provides a very useful tool for quantifying the structure of fractals.

Two other properties of the Cantor set are worth noting, although they are not fractal properties as such: *C has measure zero* and *it consists of uncountably many points.* These properties are clarified in the examples below.

EXAMPLE 12.3. Show that the **measure** of the Cantor set is zero, in the sense that it can be covered by intervals whose total length is arbitrary small.

Solution: Fig. 12.2 shows that each set S_n completely covers all the sets that come after it in the construction. Hence, the Cantor set $C = S_\infty$ is covered by *each* of the sets S_n. So the total length of the Cantor set must be less than the total length of S_n, for any n. Let L_n denote the length of S_n. Then from Fig. 12.2 we see that $L_0 = 1$, $L_1 = \dfrac{2}{3}$,

$L_2 = (\frac{2}{3})(\frac{2}{3}) = (\frac{2}{3})^2$, and in general, $L_n = (\frac{2}{3})^n$. Since $L_n \to 0$ as $n \to \infty$, the Cantor set has a total length of zero.

Example 12.3 suggests that the Cantor set is "small" in some sense. On the other hand, it contains tremendously many points – uncountably many, in fact.

EXAMPLE 12.4. Show that the Cantor set C is uncountable.

Solution: First we prove that the Cantor set has an elegant characterization: it consists of all points $c \in [0, 1]$ that have no 1's in their base-3 expansion.

The idea of expanding numbers in different bases may be unfamiliar. Let's see how to write an arbitrary number $x \in [0, 1]$ in base-3. We expand in powers of $\frac{1}{3}$: thus if $x = \frac{a_1}{3} + \frac{a_2}{3} + \frac{a_3}{3} + \cdots$, then $x = 0.a_1\, a_2\, a_3 \cdots$ in base-3, where the digits a_n are 0, 1 or 2. This expansion has a nice geometric interpretation (Fig. 12.3)

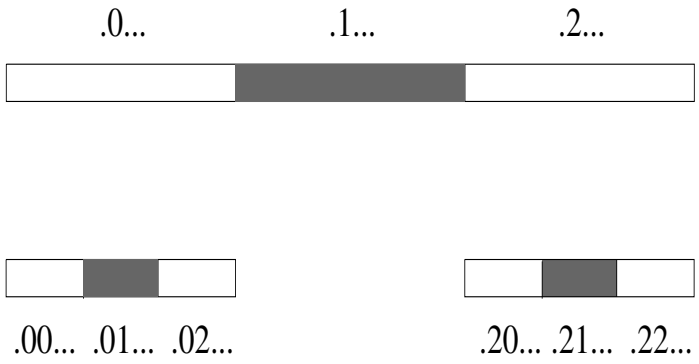


Fig. 12.3 Base-3 numbers.

It we imagine that $[0, 1]$ is divided into three pieces, then the first digit a_1 tells us whether x is in the left, middle, or right piece. For instance, all numbers with $a_1 = 0$ are in the left piece. (Ordinary base-10 works the same way, except that we divide $[0, 1]$ into ten pieces instead of three.) The second digit a_2 provides more refined information: it tells us whether x is in the left, middle, or right third of a given piece. For instance, points of the form $x = .01\cdots$ are in the middle part of the left third of $[0, 1]$, as shown in Fig. 12.3.

Now think about the base-3 expansion of points in the Cantor set C. We deleted the middle third of $[0, 1]$ at the first stage of constructing C; this removed all points whose first

digit is 1. So those points can't be in C. The points left over (the only ones with a chance of ultimately being in C) must have 0 or 2 as their first digit. Similarly, points whose *second* digit is 1 were deleted at the next stage in the construction. By repeating this argument, we see that C consists of all points whose base-3 expansion contains no 1's, as claimed.

What about endpoints like $\frac{1}{3} = .1000...$? It's in the Cantor set, yet it has a 1 in its base-3 expansion. Does this contradict what we said above? No, because this point can also be written solely in terms of 0's and 2's, as follows: $\frac{1}{3} = .1000... = .0222...$. By this trick, each point in the Cantor set can be written such that no 1's appear in its base-3 expansion, as claimed.

Now we use the above result to prove that the Cantor set is uncountable. This is just a rewrite of the Cantor diagonal argument of Example 12.2. Suppose there were a list $\{c_1, c_2, c_3, \cdots\}$ of all points in C. To show that C is uncountable, we produce a point $\bar{c}$ that is in C but not on the list. Let c_{ij} denote the jth digit in the base-3 expansion of c_i. Define $\bar{c} = \bar{c}_{11}\,\bar{c}_{22}\cdots$, where the overbar means we switch 0's and 2's: thus $\bar{c}_{nn} = 0$ if $c_{nn} = 2$ and $\bar{c}_{nn} = 2$ if $c_{nn} = 0$. Then $\bar{c}$ is in C, since it's written solely with 0's and 2's, but $\bar{c}$ is not on the list, since it differs from c_n in the nth digit. This contracts the original assumption that the list is complete. Hence, C is uncountable.

12.2.2 Koch curve. This example was originally described by von Koch in 1904. If the Cantor set can be considered a process of removing smaller and smaller segments from a set of points initially on line, then Koch curve is the complement operation of adding finer and finer length structure to an initial line segment.

We begin with a geometric construction that starts with a straight line segment S_0. To generate S_1, we delete the middle third of S_0 and replace it with the other two sides of an equilateral triangle. Subsequent stages are generated recursively by the same rule: S_n is obtained by replacing the middle third of each line segment in S_{n-1} by the other two sides of an equilateral triangle. The limiting set $K = S_\infty$ is the von Koch curve.

What is the dimension of the von Koch curve? Since it's a curve, you might be tempted to say it's one-dimensional. But the trouble is that K has *infinite arc length*! To see this, observe that if the length of S_0 is L_0, then the length of S_1 is $L_1 = \frac{4}{3}L_0$. The length

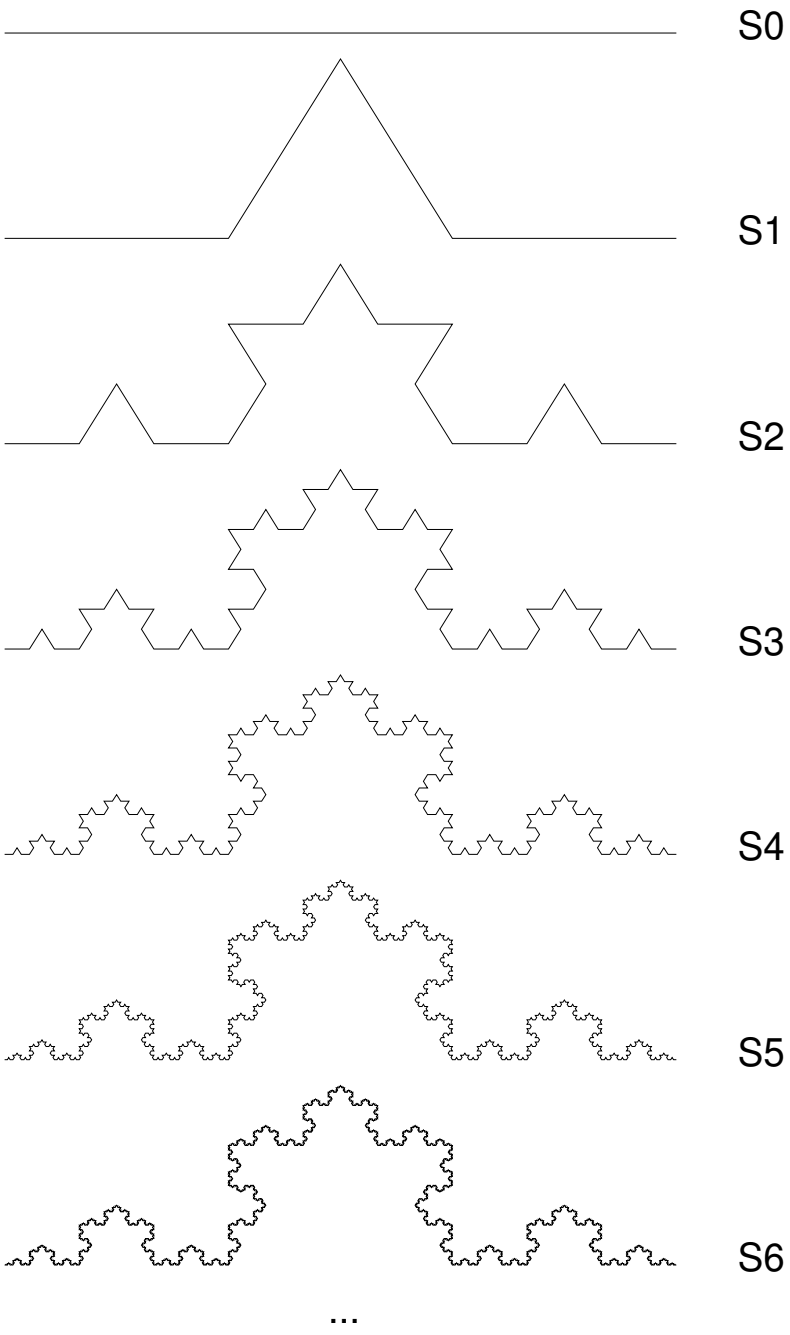


Fig. 12.4 von Koch curve K.

increases by a factor $\frac{4}{3}$ at each stage of the construction, so $L_n = (\frac{4}{3})^n L_0 \to \infty$ as $n \to \infty$.

Moreover, the arc length between *any* two points on K is infinite, by similar reasoning. Hence, points on K aren't determined by their arc length from a particular point, because every point is infinitely far from every other!

This suggests that K is more than one-dimensional. But would we really want to say the K is two-dimensional? It certainly doesn't seem to have any "area." So the dimension should be *between* 1 and 2, whatever that means.

12.2.3 Fractal dimension. Thus far we have two examples of fractal sets but do not have any test to determine if a set of points is fractal. There are many measures of the dimension of a set of points. we will describe a very intuitive or geometric definition called the *capacity* or *box-counting* dimension. Other definitions, which incorporate deeper mathematical subtleties, may be found in Mandelbrot (1977), Farmer et al. (1983), or Feder (1988) as well as in the next section. We begin with the measurement of the dimension of points along a line or distributed on some area.

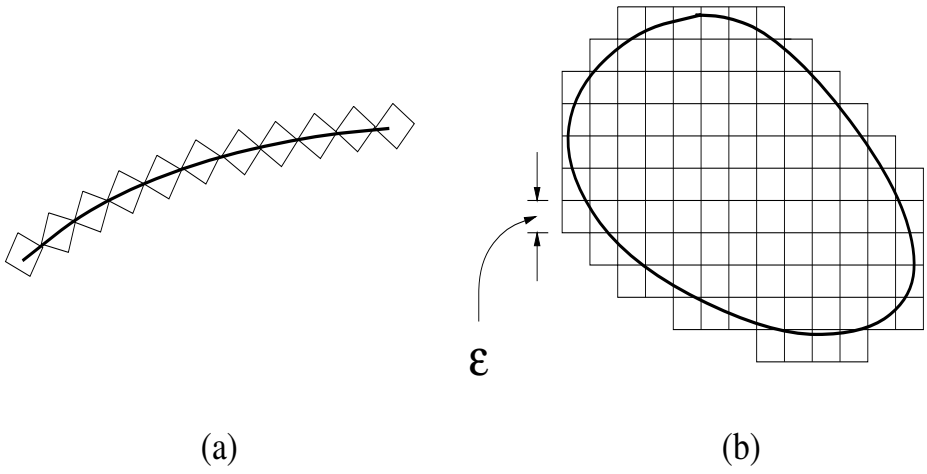


Fig. 12.5 Capacity dimension: (a) $N(\epsilon) \propto \frac{L}{\epsilon}$; (b) $N(\epsilon) \propto \frac{A}{\epsilon^2}$.

First consider a *uniform* distribution of N_0 points along some line or one-dimensional manifold in a three-dimensional space, as shown in Fig. 12.5(a). We then ask how we can *cover* this set of points with small cubes with sides of length ϵ. (One can also use spheres of radius ϵ). To be more specific, we calculate the minimum number of such cubes $N(\epsilon)$ to cover the set ($N(\epsilon) < N_0$). When N_0 is large and ϵ small enough, the number of cubes to cover a line will be

$$N(\epsilon) \propto \frac{L}{\epsilon}$$

where L is the measure of length of a line.

Similarly, if we distribute points uniformly on some two-dimensional surface in three-dimensional space, one will find that the minimum number of cubes to cover the set will be (Fig. 12.5(b)):

$$N(\epsilon) \propto \frac{A}{\epsilon^2}$$

where A represents the area bounded by the smooth curve.

Then, in general, it is natural to define the dimension by the following **power law**:

$$N(\epsilon) \propto \frac{1}{\epsilon^d}\,. \tag{12.1}$$

This power law also holds for most fractal sets, except that d is no longer an integer. By analogy with the classical case, we interpret d as a dimension, usually called the **capacity** or **box dimension**. An equivalent definition can be obtained by taking the logarithm of both sides of equation (12.1) as

$$d_C = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)} \tag{12.2}$$

if the limit exists, where the subscript C denote the capacity dimension. Implicit in this definition is the requirement that the number of points in the set be large or $N_0 \to \infty$.

A set of points is said to be fractal if its dimension is noninteger – hence the term *fractal dimension.*

In the two examples of the Cantor set and Koch curve, the fractal dimension can be calculated exactly.

EXAMPLE 12.5. Find the capacity dimensions for (a) the Cantor set; and (b) the Koch curve.

Solution:

(a) Recall that the Cantor set is covered by each of the sets S_n used in its construction (Fig. 12.2). Each S_n consists of 2^n intervals of length $(\frac{1}{3})^n$, so if we pick $\epsilon = (\frac{1}{3})^n$, we need all 2^n of these intervals to cover the Cantor set. Hence, $N(\epsilon) = 2^n$ when $\epsilon = (\frac{1}{3})^n$. Since $\epsilon \to 0$ as $n \to \infty$, we find

$$d_C = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)} = \frac{\ln(2^n)}{\ln(3^n)} = \frac{n \ln 2}{n \ln 3} = \frac{\ln 2}{\ln 3} = 0.63092... \tag{12.3}$$

(b) Similarly, for Koch curve, we may take $N(\epsilon) = 4^n$ and $\epsilon = (\frac{1}{3})^n$, and so we have

$$d_C = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)} = \frac{\ln(4^n)}{\ln(3^n)} = \frac{\ln 4}{\ln 3} = 1.26185... \tag{12.4}$$

Remarks:

R1) These two examples illustrate a helpful trick. We used a discrete sequence $\epsilon = (\frac{1}{3})^n$ that tends to zero as $n \to \infty$, even though th definition of capacity dimension says that we should let $\epsilon \to 0$ continuously. If $\epsilon \neq (\frac{1}{3})^n$, the covering will be slightly wasteful – some cubes (boxes) hang over the edge of the set – but the limiting value of d_C is the same.

R3) These two example are self-similar and so the calculations of the fractal dimension is straightforward. The following example shows how to calculate the fractal dimension of a set which is *not* self-similar.

EXAMPLE 12.6. A square region is divided into nine equal squares, and then one of the small squares is selected at random and discarded. Then the process is repeated on each of the eight remaining small squares, and so on. What is the capacity dimension of the limiting set?

Solution: Fig. 12.6 shows the first two stages in a typical realization of this random construction.

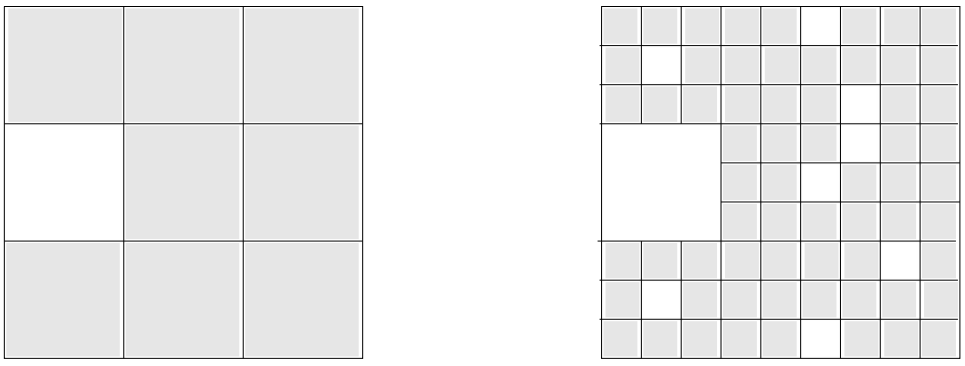


Fig. 12.6 Capacity dimension for Example 12.6.

Pick the unit of length to equal the side of the original square. Then S_1 is covered (with non wastage) by $N = 8$ squares of side $\epsilon = \frac{1}{3}$. Similarly, S_2 is covered by $N = 8^2$ squares of side $\epsilon = (\frac{1}{3})^2$. In general, $N = 8^n$ when $\epsilon = (\frac{1}{3})^n$. Hence,

$$d_C = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)} = \frac{\ln(8^n)}{\ln(3^n)} = \frac{\ln 8}{\ln 3} = 1.892789... \tag{12.5}$$

R3) It is noted that the fractal dimension of Koch curve is between 1 and 2. One way to interpret this fractal is that the distribution of points cover more than a line but less than an area. Another fractal-producing process that begins with an area distribution of points is shown in Fig. 12.7, which is called the *Sierpinski triangle* (Named after the Polish mathematician Waclaw Sierpinski, 1882-1969). At each step one removes a triangle area, creating three new triangles, but the scale is half the size of the original. One can show that this process leads to a fractal dimension of $d_C = \frac{\ln 3}{\ln 2}$.

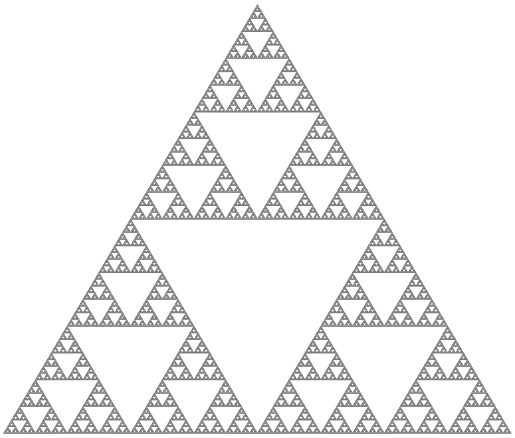


Fig. 12.7 Fractal distribution of the *Sierpinski triangle.*

12.3 Measures of Fractal Dimension

There are two criticisms of the use of capacity as a measure of fractal dimension of strange attractors – one theoretical and the other computational. First, capacity dimension is a geometric measure; that is, it does not account for the frequency with which the orbit might visit the covering cube or ball. Second, the process of counting a covering set of

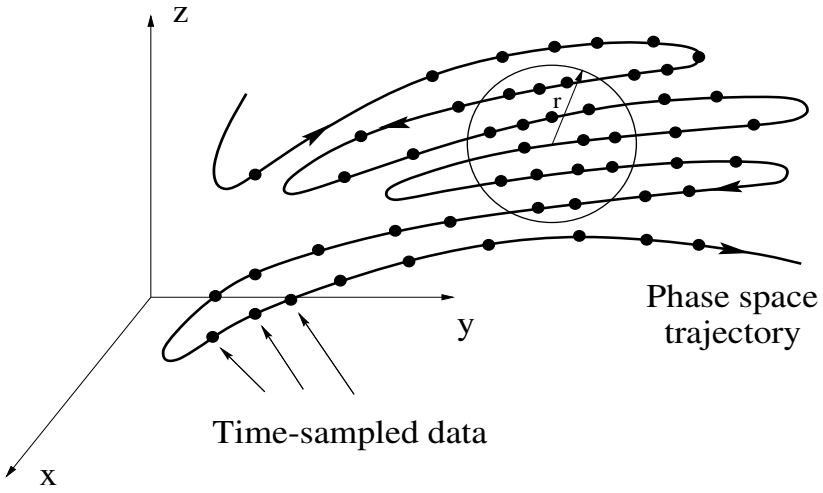


Fig. 12.8 Long time trajectory of motion in phase space.

hypercubes in phase space is very time-consuming computationally. In this section we will discuss three alternative definitions of fractal dimension which will address the shortcomings of the capacity or box-counting dimensions. However, it should be pointed out that for many strange attractors these different dimensions give roughly the same value.

12.3.1 Pointwise dimension. Let us consider a long time trajectory in phase space as shown in Fig. 12.8. first, we time sample the motion so that we have a large number of points per orbit. Second, we place a sphere or cube of radius or length r at some point on the orbit and count the number of points within the sphere $N(r)$. The probability of finding a point in this sphere is then found by dividing by the total number of points in the orbit N_0; that is,

$$P(r) = \frac{N(r)}{N_0}\,. \tag{12.6}$$

For a one-dimensional orbit, such as a closed periodic orbit, $P(r)$ will be linear in r as $r \to 0$, $N_0 \to \infty$; $P(r) \approx b\,r$. If the orbit were quasiperiodic, that is it moves on a two-dimensional toroidal surface in a three-dimensional phase space, then the probability of finding a point on the orbit in a small cube or sphere of radius r would be $P(r) \approx b\,r^2$. This leads one to define a dimension of an orbit at a point $\boldsymbol{x}_i$ (here $\boldsymbol{x}_i$ is a vector in phase space) by measuring the relative percentage of time that the orbit spends in the small sphere; that

is

$$d_P = \lim_{r\to 0} \frac{\ln P(r;\, \boldsymbol{x}_i)}{\ln r}\,. \tag{12.7}$$

For some attractors, this definition will be independent of the point $\boldsymbol{x}_i$. But for many, d_P will depend on $\boldsymbol{x}_i$ and an averaged pointwise dimension is best used. Also, for some sets of points such as a Cantor set, there will be gaps in the distribution of points so the $P(r)$ is not a smooth function of r as $r \to 0$.

To obtain an average pointwise dimension, one randomly choose a set of points $M < N_0$ and calculates $d_P(\boldsymbol{x}_i)$ at each point. The averaged pointwise dimension is given by

$$\hat{d}_P = \frac{1}{M} \sum_{i=1}^{M} d_P(\boldsymbol{x}_i)\,. \tag{12.8}$$

As an alternative, one can average the probabilities $P(r;\, \boldsymbol{x}_i)$. Choose a random *subset* of M points distributed around the attractor, where $M < N_0$. We then conjecture that

$$\lim_{r\to 0} \frac{1}{M} \sum_{i=1}^{M} P(r;\, \boldsymbol{x}_i) = a\, r^{d_P}$$

or

$$d_P = \lim_{r\to 0} \frac{\ln(1/M) \sum P(r)}{\ln r}\,. \tag{12.9}$$

In practice, if $N_0 \approx 10^3 - 10^4$ points, then $M \approx 10^2 - 10^3$.

12.3.2 Correlation dimension. As in the definition of pointwise dimension, one discretizes the orbit to a set of N points $\{\boldsymbol{x}_i\}$ in the phase space. One then calculates the distances between pairs of points, say $s_{ij} = \| \boldsymbol{x}_i - \boldsymbol{x}_j \|$, where the norm can be chosen as either the conventional Euclidean measure of distance or some equivalent measure such as the sum of absolute values of vector components. A correlation function is then defined as

$$C(r) = \lim_{N\to\infty} \frac{1}{N^2} \left(\begin{array}{l} \text{number of pairs } (i,\, j) \\ \text{with distance } s_{ij} < r \end{array} \right). \tag{12.10}$$

Fog many attractors this function has been found to exhibit a power law dependence on r as $r \to 0$; i.e.,

$$\lim_{r\to 0} C(r) = a\, r^{d} \tag{12.11}$$

so that one may define a fractal or correlation dimension using the slope of the $\ln C$ versus $\ln r$ curve (Fig. 12.9):

$$d_G = \lim_{r \to 0} \frac{\ln C(r)}{\ln r} \,. \tag{12.12}$$

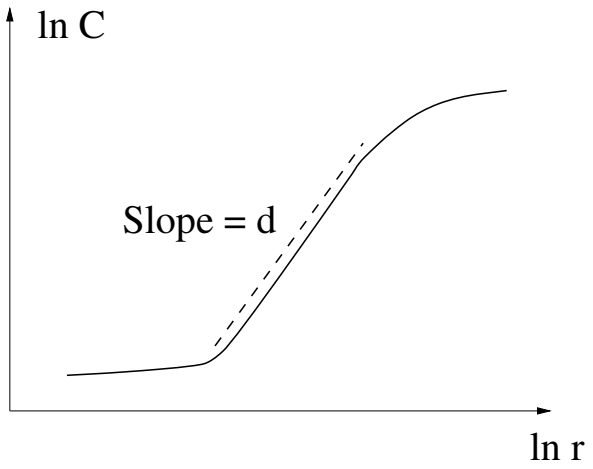


Fig. 12.9 Correlation dimension.

It has been shown that $C(r)$ may be calculated more effectively by constructing a sphere or cube at each point $\boldsymbol{x}_i$ in phase space and counting the number of points in each sphere; i.e.,

$$C(r) = \lim_{r \to 0} \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} H(r - \| \boldsymbol{x}_i - \boldsymbol{x}_j \|) \tag{12.13}$$

where $H(s) = 1$ if $s > 0$ and $H(s) = 0$ if $s < 0$. This differs from the pointwise dimension in that the sum here is performed about *every* point.

12.3.3 Information dimension. Many investigators have suggested another definition of fractal dimension that is similar to the capacity but tries to account for the frequency with which the trajectory visits each covering cube. As in the definition of capacity, one covers the set of points, whose dimension one wishes to measure, by a set of N cubes of size ϵ. This set of points is again a uniform discretization of the continuous trajectory. (It is assumed that a long enough trajectory is chosen to effectively cover the attractor whose dimension one wants to measure. For example, if the motion is quasiperiodic, the trajectory has to run long enough to "visit" all regions on the toroidal surface of the attractor.)

To calculate the information dimension, one counts the number of points N_i in each of the N cells and determines the probability of finding a point in that cell P_i, where

$$P_i \equiv \frac{N_i}{N_0} \qquad \sum^{N} P_i = 1 \tag{12.14}$$

where N_0 is the total number of points in the set. Note that $N_0 \neq N$. The *information entropy* is defined by the expression

$$I(\epsilon) = -\sum^{N} P_i \ln P_i \,. \tag{12.15}$$

(When the ln function is with respect to base 2, $I(\epsilon)$ has the units of bits.) For small ϵ it is found that I behaves as

$$I \approx d_I \ln(\frac{1}{\epsilon})$$

so that for small ϵ we may define a dimension

$$d_I = \lim_{\epsilon \to 0} \frac{I(\epsilon)}{\ln(1/\epsilon)} = \lim_{\epsilon \to 0} \frac{\sum P_i \ln P_i}{\ln \epsilon} \,. \tag{12.16}$$

To see that this definition is related to the capacity, we note that if the probabilities P_i were equal for all cells, i.e.,

$$P_i = \frac{N_i}{N_0} = \frac{1}{N} \,, \tag{12.17}$$

then

$$I = -\sum_{i=1}^{N} P_i \ln P_i = -N \frac{1}{N} \ln \frac{1}{N} = \ln N \tag{12.18}$$

so that $d_I = d_C$. In general, it can be shown that (see Farmer et al., 1983)

$$d_I \leq d_C \,. \tag{12.19}$$

Further discussions of the information dimension may be found in Farmer et al. (1983), Grassberger and Proccacia (1983), and Shaw (1984).

The information entropy is a measure of the *unpredictability* in a system. That is, for a uniform probability in each cell, $P_i = \frac{1}{N}$, I is at a maximum. If all the points are located in one cell (maximum predictability), $I = 0$ as can be seen by the calculation

$$\text{For} \quad P_i = \frac{1}{N} \,, \quad I = \ln N$$

$$\text{For} \quad P_1 = 1 \,, \quad P_i = 0 \,, \quad i \neq 1 \,, \quad I = 1 \ln 1 = 0 \,.$$

Definition (12.15) and the use of the symbol $I(\epsilon)$ are confusing in the literature. Shaw (1984) used the symbol H to denote entropy and I to denote the negative entropy $(-H)$ or *information*. Thus, for Shaw, a more predictable system (i.e., sharper P_i distribution) has *higher* information.

12.3.4 Relationship between fractal dimension and Liapunov exponents. Thus far we have defined the following fractal dimensions:

d_C the capacity (12.2),

d_P pointwise dimension (12.7),

d_G correlation dimension (12.12),

d_I information dimension (12.16).

Grassberger and Proccacia (1983) have shown that the information dimension and the correlation dimension are lower bounds on the capacity dimension; i.e.,

$$d_G \leq d_I \leq d_C \,. \tag{12.20}$$

For many of the standard strange attractor, however, all three were very close (see Table 12.1).

I summary, on can say that the capacity dimension takes no account of the distribution of points between covering cells, whereas the information entropy dimension measures the probability of finding a point in a cell. Finally, the correlation dimension accounts for the probability of finding two points in the same cell. A further relationship between fractal dimension, information entropy, and Liapunov exponents was made by Kaplan and Yorke (1978).

Table 12.1 Fractal dimension of selected dynamical systems.

Name of systems	Dimension	Type	Source of data
Hénon map ($\alpha = 1.4,\ \beta = 0.3$)	1.26 1.21 ± 0.01	Capacity correlation	Grassberger and Proccacia (1983)
Logistic map ($\lambda = 3.5699456$)	0.538 0.500 ± 0.005	Capacity correlation	Grassberger and Proccacia (1983)
Lorenz equations	2.06 ± 0.01 2.05 ± 0.01	Capacity correlation	Grassberger and Proccacia (1983)
Two-well potential ($f = 0.16,\ \omega = 0.8333$)	2.14 ($\gamma = 0.15$) 2.61 ($\gamma = 0.06$)	Correlation	Moon and Li (1985)
Chua's circuit	2.82	Liapunov	Matsumoto et al. (1985)